Some Aspects of Longitudinal Data Analysis

Peter J. Ricci

Thesis Submitted for the Degree of Doctor of Philosophy,

Department of Statistics, The University of Adelaide,

April 26, 1994

Awarded 1994

Contents

Errata

1. Page iv, line 3, change '... great deal of work for ...' to '... great deal of work done for ...'.

2. Page 16, line 1, change '... are also discussed which may permit modelling ...' to '... are also discussed. The extended quasi-likelihood equations allow modelling ...'.

3. Page 18, line 5, change '... redress the bias nature ...' to '... redress the biased nature'.

4. Page 27, the line under $E(\boldsymbol{Y}) = \sum_{i=1}^{k} \boldsymbol{T}_i \boldsymbol{\Theta} \boldsymbol{M}_i'$, change '... accordingly and is the sum ...' to ' ... accordingly. This is the sum ...'.

5. Page 38, line 1, change '... using likelihood methods (a), (b) and (c).' to '... using likelihood methods ML, PREML and REML.'.

6. Page 53, line 18, change '... WLS method, technically it a GLS ..' to '... WLS method, technically it is a GLS ...'.

7. Page 58, line 19, replace 'The modified ... Table 4.2.' by 'The Wald test statistics Q_{ij} for the modified sequential approach are given in Table 4.2.'

8. Page 59, add footnote to Table 4.2, 'The Wald statistics are approx. distributed as χ_1^2'.

9. Page 81, line 2, add the citation number [166] to the reference Zeger (1988).

10. Page 97, add the following sentences to the end of the first paragraph, 'The Base parameter β_1 signifies no change over time. Parameter β_{21} corresponds to a separate time effect for Trt at the first time point and β_2 indicates the next three time points are invariant over time. Similarly for Base.Trt and Age.'

11. Page 118, line 11, change '... distributed, I suggest a 'working covariance'...' to '... distributed, a 'working covariance'...'.

12. Page 121, line 5, remove the 'and' in '..., and which ...'.

13. Page 131, lines 19 and 20, change '... property (Model 1C instead of 1B) was the only one to converge.' to '... property (Model 1C instead of 1B was the only one to converge).'

14. Page 151, line 10, change '..., I suggest using a second order GEE for ϕ with the ...' to '..., a second order GEE for ϕ could be used with the ...'.

15. Page 156, line 5, change 'Consequently I consider Model 1 as the ...' to 'Consequently Model 1 is considered the ...'.

Summary

The analysis of longitudinal data is a very important problem in a wide range of statistical fields. There has been a great deal of work for continuous (Normal and non-Normal) and discrete data, involving parametric and semi-parametric approaches. In this thesis both the Normal and non-Normal cases are considered and the sum of profiles idea of the growth curve model (Normal theory) is extended to non-Normal data.

Chapter 1 provides a literature review of the area and a more detailed description of the thesis.

Chapter 2 introduces the profile model (growth curve model) in a canonical form. The straightforward extension to the sum of profiles model is discussed. We also describe residual maximum likelihood.

In chapter 3 estimation of the parameters for the profile model is considered under maximum likelihood and residual maximum likelihood. Two versions of residual maximum likelihood are derived and used.

Repeated categorical responses are considered in chapter 4. Extensions to the Stanek and Diehl (1988) [134] linear model are given. Connections with the sum of profiles model are discussed.

The theory of quasi-likelihood and generalized estimating equations is introduced in chapter 5. Second order extensions of the generalized estimating equations are also presented. Estimating equations obtained from a class of quadratic models are also considered.

Time dependency of the mean (and dispersion parameters) is one of the topics ex-

amined in chapter 6. The dispersion parameters are also considered as functions of covariates and are estimated using another set of generalized estimating equations.

Application of the GEE methodology of chapters 5 and 6 to three examples is the focus of chapter 7. Positive and negative aspects of the methods are discussed.

In chapter 8 a data set is examined where generalized estimating equations provide a natural approach for analysis. A number of special nested correlation structures are proposed for this data set and tests for these structures are considered.

A Signed Statement

I declare that this thesis contains no material which has been accepted for the award of any other degree or diploma in any University. To the best of my knowledge and belief, this thesis contains no material previously published or written by any other person, except where due reference is made in the text of the thesis. I consent to the thesis being made available for photocopying and loan if it is accepted for the award of the degree.

Peter J. Ricci

Acknowledgements

I sincerely thank my supervisors Dr. Ari Verbyla and Dr. Bill Venables for their patience, encouragement and guidance. They have given me many hours of their time and provided invaluable knowledge and assistance during the formation of this thesis.

To the other staff of the Statistics Department, I offer thanks for the active interest they have taken in my studies. I also thank DSTO for providing me with financial support during my candidature. I thank my parents for giving me the opportunity to study, and my family and friends for providing support and encouragement.

Finally I thank my dear wife Carolyn for the love and caring support that she offered through this difficult period of our lives and for proofreading this thesis.

Chapter 1

Introduction

1.1 Background

1.1.1 The Profile and Sum of Profiles Model

The analysis of repeated measures or longitudinal data has been examined extensively in the literature for both Normal and non-Normal cases. The common feature in all cases is that repeated measurements within each data unit cannot be presumed to be mutually independent, requiring a more elaborate estimation and analysis procedure.

A method for the analysis of multivariate Normal data is sometimes referred to as profile analysis. It is usually associated with the analysis of growth curves where observations occur at specific time points, usually evenly spaced. Interest focuses on estimation of the mean growth curves and their associated confidence bands, or tests of hypotheses concerning treatment effects. Profile analysis may also be used to analyse spatial data or situations where several similar features have been measured at a single time point.

Profile analysis involves a two-stage specification of the mean response profiles. Defining the form of the repeated measurements for the individual units represents one stage, such as by q^{th}-order polynomial representation. The other stage is achieved by specifying the effect of treatments on the individual profiles.

Wishart (1938) [161] was apparently responsible for the first work on profile analysis. His analysis involved fitting second order orthogonal polynomials to repeated measurements of bacon pigs. A separate analysis of variance was conducted for the estimated linear and quadratic orthogonal polynomial coefficients. Rowell and Walters (1976) [122] followed this approach but considered higher order polynomials as well. Box (1950) [13] used an analysis of variance approach based on differences and discussed a multivariate analysis of variance, taking initial measurements as covariates. Leech and Healy (1959) [77] also extended Wishart's (1938) [161] approach, focusing on growth rates. They took differences and also suggested using the initial measurement as a covariate.

Rao (1959) [112] used a multivariate model, where the mean growth curve for a single group of units is modelled as a polynomial. Coefficients are estimated by weighted least squares, and a model test of fit and confidence intervals are derived. Elston and Grizzle (1962) [37] considered three models, one of which is Rao's multivariate model. Another model, a mixed model where the coefficients of individual polynomials follow independent Normal distributions, is modified by Elston (1964) [36] to allow correlated coefficients, resulting in a centrosymmetric (Aitken (1956) [2] (p. 124)) covariance matrix. An alternative to polynomial growth curve models recommended by Hills (1968) [57] is to use repeated differences.

Potthoff and Roy (1964) [108] introduced a generalized multivariate linear model (GMANOVA or growth curve model) which allows simultaneous modelling of the profile with complex experimental designs. Their analysis however, depended on an arbitrary matrix. Rao (1965, 1966, 1967) [113, 114, 115] removed this arbitrary nature by stating that the correct approach is to base inferences on a conditional model. The conditional model is used by Khatri (1966) [63] who derives the explicit maximum likelihood estimates as well as test procedures and confidence intervals.

Grizzle and Allen (1969) [51] presented some distributional results and discussed the problem of loss of efficiency in estimation when a large number of covariates must be used in the conditional approach. Removal of covariates on the basis of low correlations

introduces non-unique inferences; Verbyla (1986) [146] showed that the only way to overcome this is to model the covariance structure. Lee (1974) [76] and Baksalary, Corsten and Kala (1978) [7] also discussed the relationship between the conditional approach and that of Potthoff and Roy (1964) [108]. Fujikoshi and Rao (1991) [47] discussed two types of formulation for testing hypotheses of redundancy for a *given* set of conditioning covariables; unfortunately in practice the analyst must decide which set to use.

Sugiura and Kubokawa (1988) [139] considered estimation of growth curve models with a common matrix of unknown mean parameters, but differing covariance matrices for two-sample and k-sample cases. A two-stage estimation procedure for the parameters in the growth curve model is given in Kubokawa (1989) [68] such that the covariance matrix is bounded above by a preassigned matrix, and a fixed-width confidence region is proposed. Kubokawa, Saleh and Morita (1992) [69] consider estimating the mean parameter matrix under a set of quadratic loss functions.

Lee (1991) [75] focuses on the selection of models for the growth curve model with regard to the covariance matrix. Likelihood ratio tests and selection procedures based on sample reuse as well as predictions are developed. The influence or effect of deleting measurements on growth curve estimates is studied in Liski (1991) [88], and a proposed measure for detecting influential versus non-influential measurements is discussed. Von Rosen (1991) [154] derived the moments of the maximum likelihood estimators of GMANOVA.

Žežula (1993) [170] examines a growth curve model where the experimental units are not necessarily independent. The covariance structure considered follows a simpler form than the kronecker product of two unknown matrices that arises because of the possible between-unit dependency. An estimate of the value of a parametric function of the vector of covariance components is derived. If the between individual covariance structure is completely unknown then a uniformly best estimator of the corresponding variance matrix is given.

Reinsel (1982) [116] extended the GMANOVA of Potthoff and Roy (1964) [108] to a multivariate random effects regression model. Estimation of an individual's random coefficients and prediction of future responses given values of previous responses for that individual is considered in Reinsel (1984) [117] in the multivariate context. Chinchilli and Carter (1984) [23] derive a likelihood ratio test for the patterned covariance matrix considered in Reinsel (1982) [116].

Generalizations which permit missing data have been considered by Kleinbaum (1973) [64], Srivastava and McDonald (1974) [132] (nested patterns of missing data), Woolson, Leeper and Clarke (1978) [164] and Woolson and Leeper (1980) [163]. Srivastava (1985) [133] explicitly derived likelihood ratio tests for Kleinbaum's (1973) [64] growth curve problem. Jennrich and Schluchter (1986) [59] also considered the analysis of unbalanced or incomplete repeated-measures. The problem of computing maximum likelihood estimates under a very general model for expected responses and arbitrary structural models for the within-subject covariances was examined. Tsai and Koziol (1988) [143] presented an alternative to Srivastava (1985) [133], and provided score and Wald test analogues to his likelihood ratio tests. Score and Wald test analogues to Jennrich and Schluchter's (1986) [59] likelihood ratio tests are given by Tsai and Koziol (1993) [144]. Four consistent estimators of the covariance matrix for missing data are constructed by Mensah, Elswick and Chinchilli (1993) [93] using the fact that the underlying model is GMANOVA. Comparisons are made with Kleinbaum's (1973) [64] estimator where the underlying GMANOVA model has not been used. Park, Lee and Woolson (1993) [106] develop a test procedure for distinguishing whether the missing data mechanism is missing completely at random (MCAR) or missing at random (MAR, see Rubin (1976) [124]) for incomplete Normally distributed data.

An extension to the growth curve or profile model, termed the sum of profiles model, was proposed by Verbyla and Venables (1988) [151]. Rather than modelling the treatment effects so that they all follow the same form (e.g. polynomials of same degree), the sum of profiles model permit different functional forms or profiles. The sum of profiles

model is equivalent to the analysis of Evans and Roberts (1979) [38] who modelled the treatments marginally, but only for orthogonal contrasts and complete data. Von Rosen (1989, 1990) [152, 153] extended the GMANOVA to a multivariate linear model. The extension has a sum of profiles type mean structure but involves a nesting structure of the profile design matrices. Verbyla and Cullis (1990) [149] combine the sum of profiles model and modelling the covariance matrix. The ability to handle incomplete data is demonstrated and estimation is based on residual maximum likelihood (REML).

REML was introduced by Patterson and Thompson (1971) [107] for unbalanced incomplete block designs. REML partitions the likelihood function into two parts, a marginal likelihood which is free of fixed effects and a conditional likelihood which depends on the fixed effects. Under REML, the marginal likelihood is used to estimate the covariance parameters. This contrasts with ML where the covariance parameters are estimated using the full likelihood; this makes no allowance for the loss in degrees of freedom in estimating the fixed effects. To find estimates of the fixed effects we use the conditional likelihood. Harville (1977) [55] described the properties of ML and REML along with their strengths and deficiencies; see also Cooper and Thompson (1977) [25]. Verbyla (1990) [148] presented an alternative derivation of the likelihood to that of Harville (1974) [53] and Cooper and Thompson (1977) [25], resulting in a useful formulation of residual maximum likelihood.

A number of articles have appeared recently that use REML in conjunction with longitudinal data. A discussion on how various linear patterned covariance structures affect the estimated variance of mean parameters estimated for balanced longitudinal data was presented by Lange and Laird (1989) [72] for a family of models. Log-likelihood functions for the models considered, are given under maximum likelihood and REML for the variance parameters. Explicit relations between these estimators from models within the family are derived. A modified version of REML was applied by Lundbye-Christensen (1991) [89] to a multivariate growth curve model for pregnancy. An extension of the stochastic differential equations of Sandland and McGilchrist (1979) [125]

to model growth data for designed experiments was given in Cullis and McGilchrist (1990) [27]. Parameters are estimated by REML, and extensions to incomplete data are presented.

Laird and Ware (1982) [71] considered a family of two-stage models (based on Harville (1977) [55]) for longitudinal data, examining both maximum likelihood and Bayes estimation. The Bayesian approach was considered by Harville (1974, 1976) [53, 54] and Dempster, Rubin and Tsutakawa (1981) [32], and the Bayesian formulation of the models yields REML estimates of the variance parameters.

Chi and Reinsel (1989) [22] consider linear models that contain both random effects across units and auto-correlated within-unit errors. Such models accommodate unbalanced longitudinal data. Autocorrelation is tested by a score test and Empirical Bayes estimation is used to estimate the random effects. Lindstrom and Bates (1988, 1990) [83, 84] considered linear and nonlinear mixed effects models respectively for longitudinal data and discussed computational procedures for estimating parameters under maximum likelihood and REML. Follmann (1992) [45] examines mixed effects models for growth curves with restrictions on the random parameters. Cullis and Verbyla (1992) [28] consider non-linear and time dependent covariates, while Verbyla and Cullis (1992) [150] extend work to allow for random components arising from sampling or spatial correlations as in field experiments. REML estimation is used in both articles.

1.1.2 Analysis of Non-Gaussian Longitudinal Data

There have been two major and distinct lines of development for modelling longitudinal data with a categorical response. The first and earliest is reviewed by Koch *et al.* (1977) [66]. This approach is applied to complete split-plot design data and longitudinal analysis of variance designs and is an extension of the modelling of the multivariate categorical responses described by Grizzle *et al.* (1969) [52].

Suppose that there are d conditions under which the categorical response (c categories) variable is measured. The d conditions are usually the time points at which

observations are taken. Then $r = c^d$ represents the number of possible multivariate response profiles. Similarly it is presumed that we have s subpopulations or treatments and the data may be summarized by an $s \times r$ contingency table. The joint probability of the frequencies is the product multinomial model. Standard likelihood methods would require specification of the multivariate distribution of the response profiles, but Koch *et al.* (1977) [66] argued that inferences on covariate and marginal time effects may be based on first-order marginal distributions of the profiles. Differentiable functions of the marginal probabilities are modelled as linear models of covariates of interest and parameters are estimated by weighted least squares. A consistent estimator of the covariance matrix for large samples is given and an asymptotic goodness-of-fit test and tests of linear contrasts of parameters are provided.

Koch *et al.* (1972) [65] extended the general linear approach of Grizzle *et al.* (1969) [52] for categorical data analysis to incomplete response data. Other authors who apply Grizzle *et al.* (1969) [52] methods for correlated categorical data include Woolson and Clarke (1984) [162], where modifications are made allowing incomplete response data by adding additional outcome categories designating survey-specific missing data. A test of the missing data mechanism for repeated categorical data incorporated into the Grizzle *et al.* (1969) [52] method is given by Park and Davis (1993) [105].

Beitler and Landis (1985) [8] applied a two-way analysis of variance model for multinomial responses. The variance components are estimated by a reduction sum of squares method giving an estimate of the covariance matrix of the observed response proportions. The covariance matrix is then inserted into the weighted least squares equations of Grizzle *et al.* (1969) [52] to obtain estimates of the treatment effect parameters. The Grizzle *et al.* (1969) [52] linear model is also expanded in Bonett, Woodward and Bentler (1985) [12] to include standard log-linear models and a class of Poisson distributions. Model parameters are estimated by weighted least squares under exact and stochastic constraints.

Stanish, Gillings and Koch (1978) [135] use a ratio estimation procedure for longi-

tudinal categorical data, which is an extension of Koch *et al.* (1977) [66]. Stanek and Diehl (1988) [134] use the methods of Koch *et al.* (1977) [66], but consider the situation where the number of repeated measures is too large to permit modelling all the marginal response functions. A class of growth-curve models are developed for binary data which characterize the marginal response pattern within each response subpopulation by a subset of significant effects from the full set of polynomials. Estimation of the parameters is carried out using weighted least squares. Lipsitz, Laird and Harrington (1990) [85] discuss the finding of the design matrix for a marginal homogeneity model applied to repeated categorical data. Estimation of the parameters of the marginal homogeneity linear model follows the weighted least squares framework of Koch *et al.* (1977) [66].

The second major method for modelling longitudinal categorical data is the Generalized Estimating Equation (GEE) method, which can be thought of as an extension of quasi-likelihood theory for longitudinal data.

The quasi-likelihood method was proposed by Wedderburn (1974) [158] for estimation of the parameters in regression models, when a fully specified likelihood is not assumed because of insufficient information. For each observation, the mean and variance however is presumed to follow a particular relationship. If the unknown underlying distribution is a member of the exponential family the quasi-likelihood estimate maximizes the likelihood and so full asymptotic efficiency occurs. There is some loss of efficiency for more general models (Firth (1987) [41]).

Nelder and Pregibon (1987) [100] introduced the extended quasi-likelihood to permit formal comparison of models having different variance functions or dispersion parameters. As in the case of quasi-likelihood, the extended quasi-likelihood assumes information is available on the first two moments. Another important feature of extended quasi-likelihood is the possibility of modelling the dispersion parameter as a function of covariates of interest. Nelder and Lee (1992) [99] compared the maximum likelihood, extended quasi-likelihood and pseudolikelihood (see e.g. Carroll and Ruppert (1982) [20] or Davidian and Carroll (1987) [31]) estimators under three models. The finite sampling

properties of the estimator of the dispersion parameter for the three methods is explored by simulation. The maximum extended quasi-likelihood estimator frequently had lower mean square error than the other methods in this simulation study.

Zeger, Liang and Self (1985) [169] applied extensions to logistic regression models for longitudinal binary data and discussed an estimating equation approach. Liang and Zeger (1986) [81] presented an extension of generalized linear models for the analysis of longitudinal data. Their primary concern is modelling the dependence of the marginal distribution on covariates of interest. The dependencies among repeated measurements over time are treated as nuisance parameters. The marginal distribution is presumed to follow a generalized linear model and a 'working' correlation matrix is used. A set of generalized estimating equations lead to consistent estimates of the mean parameters even if the 'working' covariance matrix is incorrectly specified. A consistent estimate of the covariance matrix (even under incorrect specification of the 'working' correlation matrix) of the estimates of the regression parameters is presented. Under certain regularity conditions and correct specification of the mean relationship, the solution of the generalized estimating equations is asymptotic multivariate Normal, with the true parameter vector as the mean and an estimable covariance matrix. A number of possible 'working' correlation matrices are presented.

Zeger and Liang (1986) [167] extended the work of Liang and Zeger (1986) [81], the marginal distribution not being specified, but the first and second moments are presumed to follow a particular relationship. This is essentially an extension of quasi-likelihood for discrete and continuous longitudinal data. The estimating equations are referred to as generalized estimating equations (GEE) and the basic results of Liang and Zeger (1986) [81] apply.

An alternative method of estimation by maximum likelihood was proposed by Avery, Hansen and Hotz (1983) [6] for fitting multivariate probit models to longitudinal binary data. The class of estimators is termed orthogonality condition estimators and is closely related to those developed by Liang and Zeger (1986) [81]. Binder (1983) [9] addressed

the problem of finding asymptotic sampling variance estimators for finite population regression parameters in complex survey data for the class of generalized linear models. The asymptotic variance matrix obtained for the estimators is also the variance estimator described by Liang and Zeger (1986) [81].

A number of articles have appeared that consider random effects models for discrete longitudinal data. Stiratelli *et al.* (1984) [136] described a mixed-effects model for the analysis of clustered or longitudinal data, which is analogous to the Normal case discussed in Laird and Ware (1982) [71]. A subset of this model is the two-step model of Korn and Whittemore (1979) [67]. Zeger, Liang and Albert (1988) [168] considered two extensions of the generalized linear model. The first is subject-specific models for which heterogeneity is explicitly modelled, and the second is population-averaged models which describe the average response over subjects within a population. Comparisons are made between the subject-specific models and the population-averaged models for a mixed effects form and interpretation of the regression parameters for both models is discussed. GEE is used to fit both classes of models for discrete and continuous responses.

A comparison of the mixed Poisson-gamma method proposed by Thall (1988) [141] with Zeger and Liang's (1986) [167] GEE method for longitudinal interval count data was undertaken by Stukel (1993) [138]. The two methods produced similar standard errors in the examples considered except in the case of time-dependent covariates and non-Poisson-gamma data, where not surprisingly Thall's method was distinctly inferior.

Wei and Stram (1988) [159] modelled marginal distributions of discrete and continuous responses observed over equal time intervals by a quasi-likelihood function with covariates that may be time-dependent. The dependency among repeated measures is not represented by any parametric model and missing data under certain conditions are allowed. The solutions of the time-specific quasi-likelihood equations yield mean parameter estimates that are asymptotically multivariate Normal with estimable variance matrix. Multiple testing procedures for examining covariate effects over time are

given. Stram, Wei and Ware (1988) [137] described a method for comparing two groups of subjects with an ordered categorical response variable. Time-dependent covariates are considered. This is a special case of the method used by Wei and Stram (1988) [159] and the GEE approach of Liang and Zeger (1986) [81]. Separate ordinal regression models for the marginal probabilities are fitted at each time point. Tests of hypotheses of differences in the marginal distribution of the responses between groups are discussed.

Damaraju (1993) [29] considers the Stram, Wei and Ware (1988) [137] semi-parametric approach (time-specific proportional odds models) but allows for non-proportional odds at each time point. In this case the covariate effects vary across the categories of the ordinal response variable.

Ware, Lipsitz and Speizer (1988) [157] reviewed the marginal models of Koch *et al.* (1977) [66], Liang and Zeger (1986) [81] and Stram, Wei and Ware (1988) [137] and discussed the relative strengths of these approaches. Connections are made between these three approaches in the discussion by Zeger (1988) [166]. Agresti (1989) [1] also reviewed marginal models for repeated observations on categorical response variables including the multinomial model of Koch *et al.* (1977) [66] and the GEE approach of Liang and Zeger (1986) [81].

Lefkopoulou, Moore and Ryan (1989) [78], Laor and Cohen (1992) [73] and Cologne *et al.* (1993) [24] applied GEE to specific data problems. Generalized Wald and score test statistics are described by Rotnitzky and Jewell (1990) [121] for the regression parameters of the GEE for cluster correlated data. Lefkopoulou and Ryan (1993) [79] derive a class of quasi-likelihood score tests for multiple binary outcomes. Special cases of this class are shown to correspond to tests proposed by other authors. Extensions to allow for clustered data are discussed and the multivariate models examined in Lefkopoulou, Moore and Ryan (1989) [78] are considered. Clustered outcomes and the GEE approach are also examined by Catalano and Ryan (1992) [21].

The work by Miller and Landis (1991) [95], who considered generalized variance component models for clustered categorical responses, is adapted by Von Tress (1993)

[155] to longitudinal data with polytomous responses. Each response has a multinomial distribution, and estimation is by GEE. An identity link is considered. Miller, Davis and Landis (1993) [94] also extend the GEE approach to polytomous response variables. Under certain assumptions, the GEE's are shown to reduce to the weighted least squares equations discussed in Koch *et al.* (1977) [66].

Thall and Vail (1990) [142] described an extension of the GEE for longitudinal count data which provides for estimation and inference on the covariate effects on the mean, and for the variance and covariance parameters. Random subject and time effects are incorporated, leading to a class of covariance models. These covariance models can characterize heteroscedasticity, over-dispersion and dependency among repeated observations. Parameters are estimated by alternating between the GEE for the mean parameters and a set of moment equations for subject and time variance components. The combined estimates of mean and variance parameters is shown to be consistent and asymptotically Normal.

Carr and Chi (1992) [19] described analysis of variance type models for longitudinal data, and if the data is complete, then under these models the solutions to the GEE are simplified and of closed-form. Hypothesis testing is also discussed. Smith and Hadgu (1992) [130] used the GEE approach to obtain estimates of sensitivity and specificity when the data consists of correlated binary responses. Lipsitz, Laird and Harrington (1992) [87] considered a three-stage generalized least squares method for longitudinal binary data, which is similar to the methods of Ware (1985) [156] and Jennrich and Schluchter (1986) [59]. The estimates are asymptotically equivalent to the estimates obtained from GEE. O'Hara Hines and Lawless (1993) [102] examine robust asymptotic covariance matrix estimators for the regression parameters in toxicological mortality data grouped over time. These estimators are similar to those discussed in Liang and Zeger (1986) [81] and Zeger and Liang (1986) [167]. Wypij, Pugh and Ware (1993) [165] discuss the use of regression splines and GEE in the modelling of pulmonary function growth.

Prentice (1988) [109] outlined many of the important approaches in modelling correlated binary data with a corresponding vector of covariates at each outcome time. Extensions of the GEE are also proposed which allow joint inference on the regression parameters and pairwise correlations among observations. Lipsitz, Laird and Harrington (1991) [86] suggested using the odds ratio as an alternative to the pairwise correlation as a measure of association. The estimating equations of Prentice (1988) [109] are modified appropriately. Paik (1992) [104] considered an extension of the GEE that allows for heterogeneity of the dispersion parameter. *This is equivalent to the extension considered in chapter 6 and which was derived independently (Ricci and Verbyla (1991) [118]).*

A family of multivariate models that are parameterized in terms of marginal means and pairwise correlations was proposed by Zhao and Prentice (1990) [171] for correlated binary data. Score equations and Fisher's Information matrix are simplified by expressing the models in terms of convenient underlying parameters. When the number of responses is small, estimation by maximum likelihood is proposed for a special case of the multivariate model. Specifying a 'working' covariance matrix instead leads to a system of generalized estimating equations and solutions related to Liang and Zeger (1986) [81]. Consistency and the asymptotic distribution of the estimators are shown informally. Prentice and Zhao (1991) [110] extended the arguments of Zhao and Prentice (1990) [171] to the more general case of multivariate discrete and continuous responses. The models discussed in Darlington and Farewell (1992) [30] are a special case of the formulation given in Prentice (1988) [109] and Zhao and Prentice (1990) [171] but allow direct likelihood based inference procedures in contrast to the use of GEE. Zhao, Prentice and Self (1992) [172] noted that the GEE arises as maximum likelihood equations under any special case of the multivariate model introduced in Zhao and Prentice (1990) [171]. Related to the pseudo-maximum likelihood approach of Zhao and Prentice (1990) [171], Fitzmaurice and Laird (1993) [43] considered analysis of correlation binary responses but the associations between responses are modelled in terms of log-odds ratios, rather than correlations.

An interesting review of analytical methods for regression models for longitudinal categorical data is presented by Fitzmaurice, Laird and Rotnitzky (1993) [44]. Comparisons are made between likelihood-based approaches and the GEE approach. Useful discussions follow, by a number of authors including Zeger, Liang, Prentice and McCullagh.

Liang, Zeger and Qaqish (1992) [82] discussed a class of models for the marginal expectations of each response and for pairwise association. The marginal models are compared with log-linear models and pairwise association is in terms of the odds ratio. Two GEE approaches are compared for parameter estimation. The first approach focuses on the regression parameter as in Liang and Zeger (1986) [81], Prentice (1988) [109] and Lipsitz, Laird and Harrington (1991) [86] and is referred to as GEE1. The second approach simultaneously estimates the regression and association parameters, is referred to as GEE2 and was introduced in Zhao and Prentice (1990) [171]. Qaqish and Liang (1992) [111] considered a model for correlated binary data where the marginal probabilities and odds ratios may have regression structures that include multiple classes and multiple levels of nesting. Estimation is based on the second-order generalized estimating equations GEE2.

An alternative to GEE1 for multivariate binary data is presented in Carey, Zeger and Diggle (1993) [18]. The association amongst random variables in terms of the odds ratios is of interest. The algorithm presented alternates iteratively between a GEE for the mean parameters under a logistic regression model and a logistic regression of each response on others from the same cluster. An offset is applied to update the pairwise odds ratio.

REML estimation has been proposed by a number of authors for generalized linear mixed models. Breslow and Clayton (1993) [15] approximate the marginal quasi-likelihood method by Laplace's method. The penalized quasi-likelihood approach leads to estimating equations for the mean parameter vector. Approximations using Normal theory REML results for the estimation of variance components are applied in an *ad*

hoc manner, yielding a set of estimating equations for the variance parameters. Drum and McCullagh (1993) [33] adapt REML to certain logistic regression models.

1.2 An Outline of the Thesis

This thesis considers methods for Normal and non-Normal data in the analysis of repeated measures. Natural extensions of Normal theory methods to the non-Normal case can lead to a simple and flexible approach in the modelling of such data.

An overview of the theory of residual maximum likelihood (REML) is presented in chapter 2. The profile model (growth curve model) of Potthoff and Roy (1964) [108] is introduced. The profile model is represented in canonical form and the maximum likelihood (ML) estimator of the mean matrix is derived. The extension of the profile model to a sum of profiles model is also considered.

Two versions of REML are considered for the profile model in chapter 3 (Ricci, Verbyla and Venables (1994) [120]). Estimation of the variance parameters of the profile model covariance matrix is presented for both versions of REML and ML. It is then shown that the two versions of REML provide more realistic estimators of the covariance matrix of the estimated mean parameters relative to ML. The reduction in bias for REML compared to ML for the standard errors becomes evident. It is demonstrated that for a number of examples considered, the estimates of the mean parameters for all three methods are either identical or very close. REML's importance in the profile context is in providing more realistic standard errors.

Chapter 4 extends the growth curve strategy of Stanek and Diehl (1988) [134] (itself based on work of Koch *et al.* (1977) [66]) for binary repeated data (Ricci and Verbyla (1991) [119]). This extension is analogous to the sum of profiles model allowing different linear models for different groups. Marginal functions of interest for response proportions are modelled in terms of polynomial matrices. The case of incomplete data is also discussed.

An overview of quasi-likelihood equations is presented in chapter 5. The extended

quasi-likelihood equations are also discussed which may permit modelling the dispersion parameter (as defined for generalized linear models) as a function of covariates of interest. A multivariate extension for the univariate dispersion case is derived in chapter 6. The extension of the quasi-likelihood equations to the GEE approach proposed by Liang and Zeger (1986) [81] and Zeger and Liang (1986) [167] is also discussed. A second order extension of the GEE is discussed which adds a second set of GEE for the correlation or odds ratio parameters. Finally a set of estimating equations obtained from a class of quadratic models for the mean and covariance parameters as proposed by Zhao and Prentice (1990) [171] and Prentice and Zhao (1991) [110] is briefly discussed. A number of 'working' covariance matrices are presented here which can be used with the estimating equations of the quadratic models leading to the second order GEE.

In chapter 6 we consider occasion-specific parameters for the mean and the dispersion in a GEE context (Ricci and Verbyla (1991) [118]). Stram, Wei and Ware (1988) [137] modelled occasion-specific parameters for the mean, but estimated the parameters separately at each time point. Zeger (1988) [166] commented that this is a solution to the GEE when we let the coefficients vary in time and use an identity (independence) working correlation matrix. How the occasion-specific regression parameters can be written in a form that allows missing values is discussed. Estimation of the mean parameters is considered with non-independent working correlation matrices. The mean model can be written in a form analogous to the sum of profiles model.

The dispersion parameters (which may also be occasion-specific) in terms of covariates of interest are modelled and another set of GEE for estimation of the covariate parameters is developed. The dispersion covariate parameters may be estimated with the mean parameters by either two separate sets of GEE or in a combined GEE for all parameters. Covariance or correlation parameters may also be estimated in their own GEE or as part of a combined GEE. Discussion of estimation by separate GEE (GEE1) or combined GEE (GEE2) is given in chapter 5. An outline of the asymptotic covariance matrix for the parameter estimates is presented in chapter 6. As acknowledged

earlier, Paik (1992) [104] independently considered dispersion models with GEE1 (but not GEE2).

Measures of the goodness of fit of a model are discussed in chapter 6.

Three examples are considered in chapter 7 for parameter estimation by the GEE approach. The first example involves occasion-specific parameters. Modelling the dispersion parameter in terms of covariates of interest is considered in two of the examples. Difficulties in estimation by GEE2 become apparent in the three examples. The positive and negative aspects of GEE1, GEE2 and moment estimation are discussed.

Chapter 8 proposes a series of nested kronecker product correlation structures for dental data for 117 rats. The GEE approach provides a natural and straightforward method for analysis of this data set. Tests of the validity of the correlation structures are given. A mixed effects model is also applied.

Chapter 2

REML and the Profile Model

2.1 Introduction

In this chapter the philosophy and theory of residual maximum likelihood (REML) is introduced. REML attempts to redress the bias nature of standard maximum likelihood (ML) techniques with respect to estimating the variance parameters.

The profile model or growth curve model is also introduced for repeated measurement data. This model has been extensively examined in the literature. A canonical representation of the profile model is considered and the maximum likelihood estimator of the mean parameters is derived. An extension of the profile model, titled the sum of profiles model is briefly described.

2.2 REML and the General Linear Model

In this section the methodology of Verbyla (1990) [148] is used to derive the REML estimators for the general linear model,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.1}$$

where $\boldsymbol{y}$ is an $n \times 1$ vector of responses, $\boldsymbol{X}$ is an $n \times p$ design matrix and is assumed to be of full column rank, and $\boldsymbol{\epsilon}$ is the random error vector which is distributed as $N(\boldsymbol{0},\ \sigma^2\boldsymbol{\Omega})$.

The matrix $\boldsymbol{\Omega}$ is assumed to be positive definite and is completely specified by an $r \times 1$ parameter vector $\boldsymbol{\gamma}$, $r < p$; if necessary we write $\boldsymbol{\Omega}(\boldsymbol{\gamma})$. We require estimates of $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\gamma}$.

The following lemma will be required:

Lemma 2.1 *Let $\boldsymbol{M}^{p\times q}$ and $\boldsymbol{N}^{p\times(p-q)}$ be matrices of full column rank q and $p-q$ respectively, such that $\boldsymbol{M}'\boldsymbol{N} = \boldsymbol{0}$. Then if $\boldsymbol{\Omega}$ is a symmetric positive definite matrix,*

$$\boldsymbol{\Omega}^{-1} = \boldsymbol{\Omega}^{-1}\boldsymbol{M}(\boldsymbol{M}'\boldsymbol{\Omega}^{-1}\boldsymbol{M})^{-1}\boldsymbol{M}'\boldsymbol{\Omega}^{-1} + \boldsymbol{N}(\boldsymbol{N}'\boldsymbol{\Omega}\boldsymbol{N})^{-1}\boldsymbol{N}'.$$

Proof: See Khatri (1966) [63].

We derive a canonical form for the general linear model (2.1). Let $\boldsymbol{L} = [\boldsymbol{L}_1^{n\times p}\ \boldsymbol{L}_2^{n\times(n-p)}]$, where $\boldsymbol{L}$ is of full column rank, and satisfies

$$\boldsymbol{L}'_1\boldsymbol{X} = \boldsymbol{I}^{p\times p} \quad \text{and} \quad \boldsymbol{L}'_2\boldsymbol{X} = \boldsymbol{0}. \tag{2.2}$$

A convenient choice for $\boldsymbol{L}_1$ is $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$. If we let $\boldsymbol{y}_1 = \boldsymbol{L}'_1\boldsymbol{y}$ and $\boldsymbol{y}_2 = \boldsymbol{L}'_2\boldsymbol{y}$, then the transformed matrix is distributed as

$$\boldsymbol{L}'\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{0} \end{bmatrix},\ \sigma^2 \begin{bmatrix} \boldsymbol{L}'_1\boldsymbol{\Omega}\boldsymbol{L}_1 & \boldsymbol{L}'_1\boldsymbol{\Omega}\boldsymbol{L}_2 \\ \boldsymbol{L}'_2\boldsymbol{\Omega}\boldsymbol{L}_1 & \boldsymbol{L}'_2\boldsymbol{\Omega}\boldsymbol{L}_2 \end{bmatrix} \right). \tag{2.3}$$

If $\boldsymbol{\Sigma}_2 = \boldsymbol{L}'_2\boldsymbol{\Omega}\boldsymbol{L}_2$, the marginal distribution is

$$\boldsymbol{y}_2 \sim N(\boldsymbol{0},\ \sigma^2\boldsymbol{\Sigma}_2),$$

which does not depend on $\boldsymbol{\beta}$.

The conditional distribution,

$$\boldsymbol{y}_1|\boldsymbol{y}_2 \sim N(\boldsymbol{\beta} + \boldsymbol{B}\boldsymbol{y}_2,\ \sigma^2\boldsymbol{\Sigma}_1) \tag{2.4}$$

where

$$\boldsymbol{B} = (\boldsymbol{L}'_1\boldsymbol{\Omega}\boldsymbol{L}_2)(\boldsymbol{L}'_2\boldsymbol{\Omega}\boldsymbol{L}_2)^{-1}$$

and

$$\begin{aligned}\Sigma_1 &= (L_1'\Omega L_1) - (L_2'\Omega L_1)(L_2'\Omega L_2)^{-1}(L_1'\Omega L_2) \\ &= (X'\Omega^{-1}X)^{-1}\end{aligned}$$

contains all the information on β. Thus y_2 acts as a covariate in (2.4) in the estimation of β.

If Ω is assumed given, then B is known and By_2 acts as an offset in (2.4). Subtracting By_2 from y_1 then leads to the simple maximum likelihood estimator of β,

$$\begin{aligned}\hat{\beta} &= [y_1 - By_2] \\ &= L_1'[I - \Omega L_2(L_2'\Omega L_2)^{-1}L_2']y \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y \qquad (2.5)\end{aligned}$$

where we have applied Lemma 2.1 and the maximum likelihood estimator is the generalized least squares estimator as expected.

The log-likelihood function for the marginal distribution of y_2 is

$$\begin{aligned}\mathcal{L}(\sigma^2,\ \gamma;\ y_2) = \ & \text{const} - \tfrac{1}{2}(n-p)\log\sigma^2 - \tfrac{1}{2}\log\{\det(L_2'\Omega L_2)\} \\ & -\tfrac{1}{2\sigma^2}(y'L_2(L_2'\Omega L_2)^{-1}L_2'y) \qquad (2.6)\end{aligned}$$

where $\det(L_2'\Omega L_2)$ is the determinant of $L_2'\Omega L_2$. The log-likelihood $\mathcal{L}(\sigma^2,\ \gamma;\ y_2)$ is the log-likelihood of the contrasts with zero mean, that is the error contrasts. Information on the error parameters is concentrated solely in the marginal distribution of y_2. This can be seen by noting that the size of the vector y_1 for the conditional distribution of y_1 given y_2 is equal to the size of the parameter vector β.

Since

$$\mathcal{L}(\beta,\ \sigma^2,\ \gamma;\ L'y) = \mathcal{L}(\beta;\ y_1|y_2)\mathcal{L}(\sigma^2,\ \gamma;\ y_2) \qquad (2.7)$$

then equating the determinants on the left and right side of (2.7) gives

$$\det(L'\Omega L) = \det(X'\Omega^{-1}X)\det(L_2'\Omega L_2)$$

and by properties of determinants

$$\det(\boldsymbol{L}_2'\boldsymbol{\Omega}\boldsymbol{L}_2) = \det(\boldsymbol{L}'\boldsymbol{L})\det(\boldsymbol{\Omega})\det(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}). \tag{2.8}$$

Hence we can then write (2.6) as

$$\begin{aligned}\mathcal{L}(\sigma^2,\ \boldsymbol{\gamma};\ \boldsymbol{y}_2) &= \text{const} - \frac{1}{2}(n-p)\log\sigma^2 - \frac{1}{2}\log\{\det(\boldsymbol{\Omega})\} \\ &\quad -\frac{1}{2}\log\{\det(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})\} - \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{2\sigma^2}\end{aligned} \tag{2.9}$$

where we have used (2.8) and

$$\boldsymbol{P} = \boldsymbol{L}_2(\boldsymbol{L}_2'\boldsymbol{\Omega}\boldsymbol{L}_2)^{-1}\boldsymbol{L}_2' = \boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}$$

by Lemma 2.1. The adjustment to the full log-likelihood is the term

$$\log\{\det(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})\}/2.$$

This is the log-likelihood as given by Harville (1974) [53] and Cooper and Thompson (1977) [25]. Harville (1974) [53] transforms the $n \times 1$ response vector $\boldsymbol{y}$ to the $p \times 1$ estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{y}$ and to a set of $n-p$ linear functions with zero means, corresponding to $\boldsymbol{L}_2'\boldsymbol{y}$ above. The likelihood is derived by the independence of these sets of linear functions. However the transformation used is dependent on $\boldsymbol{\Omega}$.

Estimates of the variance parameters σ^2 and $\boldsymbol{\gamma}$ are obtained by maximizing the log-likelihood $\mathcal{L}(\sigma^2,\ \boldsymbol{\gamma};\ \boldsymbol{y}_2)$, that is solving the score equations,

$$\frac{\partial\mathcal{L}(\sigma^2,\ \boldsymbol{\gamma};\ \boldsymbol{y}_2)}{\partial\sigma^2} = -\frac{(n-p)}{2\sigma^2} + \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{2\sigma^4} \tag{2.10}$$

$$\frac{\partial\mathcal{L}(\sigma^2,\ \boldsymbol{\gamma};\ \boldsymbol{y}_2)}{\partial\gamma_i} = -\frac{1}{2}\text{tr}\left(\boldsymbol{P}\frac{\partial\boldsymbol{\Omega}}{\partial\gamma_i}\right) + \frac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{P}\frac{\partial\boldsymbol{\Omega}}{\partial\gamma_i}\boldsymbol{P}\boldsymbol{y},\quad i = 1,\ldots,\ r, \tag{2.11}$$

where (2.10) is derived from (2.9) and (2.11) is algebraically easier to derive from (2.6). Equation (2.10) yields the REML estimator for σ^2,

$$\hat{\sigma}_R^2 = \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{n-p}. \tag{2.12}$$

Using the result

$$\frac{\partial\boldsymbol{P}}{\partial\gamma_i} = -\boldsymbol{P}\frac{\partial\boldsymbol{\Omega}}{\partial\gamma_i}\boldsymbol{P},$$

the second order derivatives are

$$\begin{aligned}\frac{\partial \mathcal{L}(\sigma^2,\ \boldsymbol{\gamma};\ \boldsymbol{y}_2)}{\partial\sigma^2\partial\sigma^2} &= \frac{(n-p)}{2\sigma^4} - \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}}{\sigma^6} \\ \frac{\partial \mathcal{L}(\sigma^2,\ \boldsymbol{\gamma};\ \boldsymbol{y}_2)}{\partial\sigma^2\partial\gamma_i} &= -\frac{1}{2\sigma^4}\boldsymbol{y}'\boldsymbol{P}\frac{\partial\boldsymbol{\Omega}}{\partial\gamma_i}\boldsymbol{P}\boldsymbol{y}\end{aligned} \tag{2.13}$$

and

$$\begin{aligned}\frac{\partial \mathcal{L}(\sigma^2,\ \boldsymbol{\gamma};\ \boldsymbol{y}_2)}{\partial\gamma_i\partial\gamma_i} &= \frac{1}{2}\mathrm{tr}\left(\boldsymbol{P}\frac{\partial\boldsymbol{\Omega}}{\partial\gamma_j}\boldsymbol{P}\frac{\partial\boldsymbol{\Omega}}{\partial\gamma_i}\right) - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{P}\frac{\partial^2\boldsymbol{\Omega}}{\partial\gamma_j\partial\gamma_i}\right) \\ &\quad -\frac{1}{\sigma^2}\boldsymbol{y}'\boldsymbol{P}\frac{\partial\boldsymbol{\Omega}}{\partial\gamma_j}\boldsymbol{P}\frac{\partial\boldsymbol{\Omega}}{\partial\gamma_i}\boldsymbol{P}\boldsymbol{y} + \frac{1}{2\sigma^2}\boldsymbol{y}'\boldsymbol{P}\frac{\partial^2\boldsymbol{\Omega}}{\partial\gamma_j\partial\gamma_i}\boldsymbol{P}\boldsymbol{y}.\end{aligned} \tag{2.14}$$

A number of results are required to evaluate the expectations of the above second order derivatives.

(a) If $\boldsymbol{D}$ is a symmetric matrix then $E(\boldsymbol{y}'\boldsymbol{D}\boldsymbol{y}) = \mathrm{tr}(\boldsymbol{\Sigma}\boldsymbol{D}) + \boldsymbol{\mu}'\boldsymbol{D}\boldsymbol{\mu}$, where $E(\boldsymbol{y}) = \boldsymbol{\mu}$ and $\mathrm{var}(\boldsymbol{y}) = \boldsymbol{\Sigma}$, and

(b) $\boldsymbol{P}\boldsymbol{\Omega}\boldsymbol{P} = \boldsymbol{P}$.

Fisher's Information matrix for $(\sigma^2,\ \boldsymbol{\gamma}')'$, represented as $\boldsymbol{\Phi}$ say, from the marginal likelihood of $\boldsymbol{y}_2$ has components,

$$\begin{array}{c} \\ \sigma^2 \\ \gamma_i \\ \gamma_j \end{array}\begin{array}{c} \begin{array}{ccc} \sigma^2 & \gamma_i & \gamma_j \end{array} \\ \begin{pmatrix} \frac{(n-p)}{2\sigma^4} & \frac{1}{2\sigma^2}\mathrm{tr}\left(\boldsymbol{P}\frac{\partial\boldsymbol{\Omega}}{\partial\gamma_i}\right) & \frac{1}{2\sigma^2}\mathrm{tr}\left(\boldsymbol{P}\frac{\partial\boldsymbol{\Omega}}{\partial\gamma_j}\right) \\ & \frac{1}{2}\mathrm{tr}\left(\left\{\boldsymbol{P}\frac{\partial\boldsymbol{\Omega}}{\partial\gamma_i}\right\}^2\right) & \frac{1}{2}\mathrm{tr}\left(\boldsymbol{P}\frac{\partial\boldsymbol{\Omega}}{\partial\gamma_j}\boldsymbol{P}\frac{\partial\boldsymbol{\Omega}}{\partial\gamma_i}\right) \\ & & \frac{1}{2}\mathrm{tr}\left(\left\{\boldsymbol{P}\frac{\partial\boldsymbol{\Omega}}{\partial\gamma_j}\right\}^2\right) \end{pmatrix} \end{array}.$$

Estimation of σ^2 and $\boldsymbol{\gamma}$ from (2.10) and (2.11) can be obtained by Fisher's method of scoring. Since an explicit estimate (2.12) of σ^2 is available, we can use (2.12) and the part of the iterative procedure that is used in estimating $\boldsymbol{\gamma}$,

$$\hat{\boldsymbol{\gamma}}_k = \hat{\boldsymbol{\gamma}}_{k-1} + \hat{\boldsymbol{\Gamma}}_{k-1}\hat{\boldsymbol{U}}_{k-1}, \quad (k\text{th iteration}), \tag{2.15}$$

where $\boldsymbol{\Gamma}$ is the submatrix formed from the bottom r rows of the inverse of $\boldsymbol{\Phi}$ and $\boldsymbol{U}$ is the score vector of $\boldsymbol{\gamma}$.

2.3 The Profile Model

Potthoff and Roy (1964) [108] introduced a generalized multivariate linear model referred to here as the profile model. The profile model is also referred to as the growth curve model or GMANOVA. As the profile model can be applied to situations other than growth curve problems, we prefer to use the more general title, due to Morrison (1967) [97].

Suppose p measurements of a response variable of interest are taken for each of n individuals or experimental units. The measurements may be taken over p time points or under p different conditions. For each individual, a $q \times 1$, $(q < p)$, vector of explanatory variables or covariates, $\boldsymbol{m}_i$, are measured at each time point. We assume that the response at the j^{th} time point, of the i^{th} individual, y_{ij}, has mean

$$E(y_{ij}) = \boldsymbol{m}_i'\boldsymbol{\theta}.$$

The vector $\boldsymbol{\theta}^{q\times 1}$ is a vector of unknown regression parameters. Consequently

$$E(\boldsymbol{y}_i) = \boldsymbol{M}\boldsymbol{\theta}$$

where $\boldsymbol{y}_i = [y_{i1},\ y_{i2}, \ldots,\ y_{ip}]'$ and $\boldsymbol{M} = [\boldsymbol{m}_1,\ \boldsymbol{m}_2, \ldots,\ \boldsymbol{m}_p]'$.

Each individual $\boldsymbol{y}_i$ is assumed to be distributed as

$$\boldsymbol{y}_i \sim N(\boldsymbol{M}\boldsymbol{\theta},\ \boldsymbol{\Omega}), \tag{2.16}$$

where $\boldsymbol{\Omega}^{p\times p}$ is an unknown covariance matrix. We will consider the case of $\boldsymbol{\Omega}$ fully parameterized as well as two parsimonious parametric forms. The linear model (2.16) can be thought of as a 'within individual model'.

If the individuals are randomly allocated to m groups or treatments then the model for the entire sample is

$$\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\Theta}\boldsymbol{M}',\ \boldsymbol{I}_n \otimes \boldsymbol{\Omega}), \tag{2.17}$$

where the rows of $\boldsymbol{Y}^{n\times p}$ correspond to the experimental units, and $\boldsymbol{X}^{n\times m}$ is a fixed design matrix specifying the treatment structure 'across individuals'. Model (2.17) permits

expectations to vary from group to group (defined for example by several factors), with the rows of $\Theta^{m\times q}$ representing the regression parameters for different treatment groups. Model (2.17) however restricts the means between groups to have the same profile design dimension and form, for example polynomial models with equal degrees. The matrices $\boldsymbol{X}$ and $\boldsymbol{M}$ are assumed to be of full column rank.

2.4 The Canonical Form

The canonical form introduced in §2.2 can be easily extended for the profile model. Let $\boldsymbol{Q}_1$, $\boldsymbol{Q}_2$, $\boldsymbol{L}_1$ and $\boldsymbol{L}_2$ be matrices of full column rank such that

$$\boldsymbol{M}'\boldsymbol{Q}_1 = \boldsymbol{I}_q, \quad \boldsymbol{M}'\boldsymbol{Q}_2 = \boldsymbol{0}, \quad \boldsymbol{L}_1'\boldsymbol{X} = \boldsymbol{I}_m, \quad \text{and} \quad \boldsymbol{L}_2'\boldsymbol{X} = \boldsymbol{0}. \tag{2.18}$$

and $\boldsymbol{Q} = [\boldsymbol{Q}_1^{p\times q}\ \boldsymbol{Q}_2^{p\times(p-q)}]$ and $\boldsymbol{L} = [\boldsymbol{L}_1^{n\times m}\ \boldsymbol{L}_2^{n\times(n-m)}]$ are non-singular. A convenient choice for $\boldsymbol{L}_1$ is $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and $\boldsymbol{L}_2$ can be chosen to have orthonormal columns, satisfying $\boldsymbol{L}_2\boldsymbol{L}_2' = \boldsymbol{I} - \boldsymbol{P}_X$ where $\boldsymbol{P}_X = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. If $\boldsymbol{Y}_{11} = \boldsymbol{L}_1'\boldsymbol{Y}\boldsymbol{Q}_1$, $\boldsymbol{Y}_{12} = \boldsymbol{L}_1'\boldsymbol{Y}\boldsymbol{Q}_2$, $\boldsymbol{Y}_{21} = \boldsymbol{L}_2'\boldsymbol{Y}\boldsymbol{Q}_1$ and $\boldsymbol{Y}_{22} = \boldsymbol{L}_2'\boldsymbol{Y}\boldsymbol{Q}_2$ denote the components of the non-singular transformation of $\boldsymbol{Y}$, $\boldsymbol{L}'\boldsymbol{Y}\boldsymbol{Q}$, we have the canonical form (similar to Gleser and Olkin (1970) [48])

$$\boldsymbol{L}'\boldsymbol{Y}\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{Y}_{11} & \boldsymbol{Y}_{12} \\ \boldsymbol{Y}_{21} & \boldsymbol{Y}_{22} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\Theta} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} (\boldsymbol{X}'\boldsymbol{X})^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{n-m} \end{bmatrix} \otimes \boldsymbol{Q}'\boldsymbol{\Omega}\boldsymbol{Q} \right),$$

$$\begin{aligned} \boldsymbol{Y}_{11}|\boldsymbol{Y}_{12} &\sim N(\boldsymbol{\Theta} + \boldsymbol{Y}_{12}\boldsymbol{B},\ (\boldsymbol{X}'\boldsymbol{X})^{-1} \otimes (\boldsymbol{M}'\boldsymbol{\Omega}^{-1}\boldsymbol{M})^{-1}), \\ \boldsymbol{Y}_{21}|\boldsymbol{Y}_{22} &\sim N(\boldsymbol{Y}_{22}\boldsymbol{B},\ \boldsymbol{I}_{n-m} \otimes (\boldsymbol{M}'\boldsymbol{\Omega}^{-1}\boldsymbol{M})^{-1}), \\ \boldsymbol{Y}_{12} &\sim N(\boldsymbol{0},\ (\boldsymbol{X}'\boldsymbol{X})^{-1} \otimes \boldsymbol{Q}_2'\boldsymbol{\Omega}\boldsymbol{Q}_2), \\ \boldsymbol{Y}_{22} &\sim N(\boldsymbol{0},\ \boldsymbol{I}_{n-m} \otimes \boldsymbol{Q}_2'\boldsymbol{\Omega}\boldsymbol{Q}_2), \end{aligned} \tag{2.19}$$

where

$$\boldsymbol{B} = (\boldsymbol{Q}_2'\boldsymbol{\Omega}\boldsymbol{Q}_2)^{-1}\boldsymbol{Q}_2'\boldsymbol{\Omega}\boldsymbol{Q}_1.$$

If $\boldsymbol{\Omega}$ is fully parameterized, the new parameters $\boldsymbol{B}$, $(\boldsymbol{M}'\boldsymbol{\Omega}^{-1}\boldsymbol{M})^{-1}$ and $\boldsymbol{Q}_2'\boldsymbol{\Omega}\boldsymbol{Q}_2$ are a one to one mapping of the $p(p+1)/2$ parameters of $\boldsymbol{\Omega}$.

2.5 Fully Parameterized Covariance Matrix

Using the canonical form (2.19), estimation of Θ is based on the conditional distribution of $\boldsymbol{Y}_{11}$ given $\boldsymbol{Y}_{12}$, analogous to estimating $\boldsymbol{\beta}$ from the conditional distribution of $\boldsymbol{y}_1$ given $\boldsymbol{y}_2$ in §2.2. Assuming Ω is known, then as occurred for $\hat{\boldsymbol{\beta}}$ in (2.5), $\boldsymbol{Y}_{12}\boldsymbol{B}$ is completely known and can be used as an offset. The ML estimator is

$$\begin{aligned}\hat{\Theta} &= \boldsymbol{Y}_{11} - \boldsymbol{Y}_{12}\boldsymbol{B} \\ &= \boldsymbol{L}_1'\boldsymbol{Y}(\boldsymbol{I} - \boldsymbol{Q}_2(\boldsymbol{Q}_2'\Omega\boldsymbol{Q}_2)^{-1}\boldsymbol{Q}_2'\Omega)\boldsymbol{Q}_1.\end{aligned}$$

Using Lemma 2.1 and (2.18), the ML estimator of Θ reduces to

$$\hat{\Theta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\Omega^{-1}\boldsymbol{M}(\boldsymbol{M}'\Omega^{-1}\boldsymbol{M})^{-1} \tag{2.20}$$

with covariance matrix

$$\text{var}(\hat{\Theta}|\boldsymbol{Y}_{12}) = \text{var}(\boldsymbol{Y}_{11}|\boldsymbol{Y}_{12}) = (\boldsymbol{X}'\boldsymbol{X})^{-1} \otimes (\boldsymbol{M}'\Omega^{-1}\boldsymbol{M})^{-1}. \tag{2.21}$$

The actual choice of $\boldsymbol{Q}_1$, $\boldsymbol{Q}_2$, $\boldsymbol{L}_1$ and $\boldsymbol{L}_2$ when Ω is known is not important. The resulting estimator (2.20) and covariance matrix (2.21) are invariant to the choice of $\boldsymbol{Q}$ and $\boldsymbol{L}$. However knowledge of Ω can provide simplifications in the estimation of Θ. If $\boldsymbol{Q}_1$ is chosen such that $\boldsymbol{Q}_2'\Omega\boldsymbol{Q}_1 = \boldsymbol{0}$, for example $\boldsymbol{Q}_1 = \Omega^{-1}\boldsymbol{M}(\boldsymbol{M}'\Omega^{-1}\boldsymbol{M})^{-1}$, then $\boldsymbol{Y}_{11}$ and $\boldsymbol{Y}_{12}$ are independent and inferences on Θ are based on the marginal distribution of $\boldsymbol{Y}_{11}$.

In practice Ω is generally unknown. In this situation Potthoff and Roy (1964) [108] proposed taking $\boldsymbol{Q}_1 = \boldsymbol{G}^{-1}\boldsymbol{M}(\boldsymbol{M}'\boldsymbol{G}^{-1}\boldsymbol{M})^{-1}$, where $\boldsymbol{G}$ is an arbitrary non-singular non-stochastic matrix. Inferences are then based on the marginal distribution of $\boldsymbol{Y}\boldsymbol{Q}_1$ (i.e. $\boldsymbol{L}'\boldsymbol{Y}\boldsymbol{Q}_1$). The problem with this approach is that using an arbitrary matrix $\boldsymbol{G}$ leads to inferences dependent on the actual choice of $\boldsymbol{G}$. Another potential problem is that the sufficiency principle may be violated because of the reduction of the data from $\boldsymbol{Y}$ to $\boldsymbol{Y}_1$.

As previously discussed in chapter 1, Rao (1965) [113] removed the arbitrary nature of the marginal analysis of Potthoff and Roy (1964) [108] by basing inferences on a

conditional model, such as those presented by the canonical forms in §2.2 and §2.4. Khatri (1966) [63] derived the explicit maximum likelihood (ML) estimators, but we will follow the approach of Grizzle and Allen (1969) [51].

2.6 The Sum of Profiles Model

In some situations the profile model (2.17) is not appropriate. A more flexible model that allows units or treatments to have different profiles is called the sum of profiles model, and was developed in Verbyla and Venables (1988) [151]. The sum of profiles model is briefly outlined using the terminology of Verbyla and Cullis (1990) [149].

Suppose that the i^{th} unit is observed at p_i of a total of p time points, permitting incomplete data. Let $\boldsymbol{X}_i$ be a $p_i \times mp$ design matrix for the i^{th} unit. $\boldsymbol{X}_i$ specifies which treatments are applied and which time points observations have been taken for the i^{th} unit. The $p_i \times 1$ response vector $\tilde{\boldsymbol{y}}_i$ is assumed to be distributed as

$$N(\boldsymbol{X}_i\boldsymbol{\beta},\ \sigma^2\boldsymbol{\Sigma}_i(\boldsymbol{\gamma})),$$

where $\boldsymbol{\beta}$ is a $mp \times 1$ parameter vector denoting time-by-treatment effects. The covariance matrices, $\boldsymbol{\Sigma}_i(\boldsymbol{\gamma})$, are presumed to follow the same form for all n units, and is dependent on the $r \times 1$ parameter vector $\boldsymbol{\gamma}$.

If we concatenate the response vector, $\boldsymbol{y} = (\tilde{\boldsymbol{y}}_1',\ \tilde{\boldsymbol{y}}_2', \ldots,\ \tilde{\boldsymbol{y}}_n')'$, then

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta},\ \sigma^2\boldsymbol{\Omega}(\boldsymbol{\gamma})), \tag{2.22}$$

where $\boldsymbol{X} = (\boldsymbol{X}_1'\ \boldsymbol{X}_2' \ldots \boldsymbol{X}_n')'$ and $\boldsymbol{\Omega} = \text{diag}(\boldsymbol{\Sigma}_i(\boldsymbol{\gamma}))$. Let $N = p_1 + p_2 + \ldots + p_n$. In most situations the $N \times mp$ design matrix $\boldsymbol{X}$ is of full column rank.

The mean parameter vector $\boldsymbol{\beta}$ is composed of m blocks of p elements. The j^{th} block represents the j^{th} treatment effect at each of the p time points. We may in some situations wish to model the blocks of $\boldsymbol{\beta}$, i.e. the treatment effects, for example by a

linear model of form,

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{M}_1\boldsymbol{\theta}_1 \\ \boldsymbol{M}_2\boldsymbol{\theta}_2 \\ \vdots \\ \boldsymbol{M}_m\boldsymbol{\theta}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{M}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{M}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{M}_m \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \\ \vdots \\ \boldsymbol{\theta}_m \end{bmatrix} = \boldsymbol{A}\boldsymbol{\theta}. \tag{2.23}$$

Model (2.23) allows different parametric forms for different blocks of $\boldsymbol{\beta}$, by different profile design matrices $\boldsymbol{M}_i^{p\times q_i}$.

When the data is complete, then we can represent (2.23) and (2.22) in matrix form, that is

$$E(\boldsymbol{Y}) = \boldsymbol{T}\boldsymbol{\Delta} = \boldsymbol{H}\begin{bmatrix} \boldsymbol{\theta}_1'\boldsymbol{M}_1' \\ \boldsymbol{\theta}_2'\boldsymbol{M}_2' \\ \vdots \\ \boldsymbol{\theta}_m'\boldsymbol{M}_m' \end{bmatrix} = \sum_{i=1}^{m} \boldsymbol{t}_i\boldsymbol{\theta}_i'\boldsymbol{M}_i \tag{2.24}$$

where $\boldsymbol{T} = [\boldsymbol{t}_1\ \boldsymbol{t}_2 \ldots\ \boldsymbol{t}_m]$ is an $n \times m$ treatment design matrix and $\boldsymbol{\Delta}$ is the $m \times p$ matrix of parameters. If all the $\boldsymbol{M}_i$ are equal then $\boldsymbol{\Delta} = \boldsymbol{\Theta}\boldsymbol{M}'$, and (2.24) is the mean of the profile model (2.17). The vectors $\boldsymbol{y}$ and $\boldsymbol{\beta}$ are formed by stacking the rows of $\boldsymbol{Y}$ and $\boldsymbol{\Delta}$ respectively and $\boldsymbol{X} = \boldsymbol{T} \otimes \boldsymbol{I}_p$. Complete data implies that $\boldsymbol{\Omega} = \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}$.

Equation (2.24) is a sum of profiles formulation. If some of the $\boldsymbol{M}_i$ are equal then we can write (2.24) as

$$E(\boldsymbol{Y}) = \sum_{i=1}^{k} \boldsymbol{T}_i\boldsymbol{\Theta}\boldsymbol{M}_i'$$

where $\boldsymbol{T}_i^{n\times m_i}$ and $\boldsymbol{\Theta}_i^{m_i\times q_i}$ are formed accordingly and is the sum of profiles form given in Verbyla and Venables (1988) [151].

If we consider (2.22), then as in §2.2, let $\boldsymbol{L} = [\boldsymbol{L}_1^{N\times mp},\ \boldsymbol{L}_2^{N\times(N-mp)}]$ be a $N \times N$ matrix of full column rank such that $\boldsymbol{L}'_1\boldsymbol{X} = \boldsymbol{I}_{mp}$ and $\boldsymbol{L}'_2\boldsymbol{X} = \mathbf{0}$. Then the results of §2.2 hold. For given $\boldsymbol{\gamma}$, the estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{y} \tag{2.25}$$

and the ML and REML estimators of $\boldsymbol{\beta}$ are found by replacing $\boldsymbol{\Omega}$ by $\hat{\boldsymbol{\Omega}}_M = \boldsymbol{\Omega}(\hat{\boldsymbol{\gamma}}_M)$ and $\hat{\boldsymbol{\Omega}}_R = \boldsymbol{\Omega}(\hat{\boldsymbol{\gamma}}_R)$ respectively. The estimators $\hat{\boldsymbol{\gamma}}_M$ and $\hat{\boldsymbol{\gamma}}_R$ are the ML and REML

estimators of γ respectively. In the case of complete data, we can write $\hat{\beta}$ as

$$\hat{\Delta} = (T'T)^{-1}T'Y$$

where the rows of $\hat{\Delta}$ are the blocks of $\hat{\beta}$.

The estimator $\hat{\beta}$ for given γ is distributed as $\hat{\beta} \sim N(\beta,\ \sigma^2\Lambda)$, where $\Lambda = (X'\Omega^{-1}X)^{-1}$. Model (2.23) indicates that the generalized least squares estimate of θ is

$$\hat{\theta} = (A'\Lambda^{-1}A)^{-1}A'\Lambda^{-1}\hat{\beta}. \tag{2.26}$$

The REML (ML) estimator is found as before, by replacing Λ by $\hat{\Lambda}_R = \Lambda(\hat{\gamma}_R)$ ($\hat{\Lambda}_M = \Lambda(\hat{\gamma}_M)$). The estimated covariance matrix of $\hat{\theta}$ is $\hat{\sigma}^2(A'\hat{\Lambda}^{-1}A)^{-1}$.

Since the matrices in (2.26) may be large, Verbyla and Cullis (1990) [149] use a procedure based on conditional regressions to simplify computations. The details of this procedure are given in Appendix C of their paper. The standard errors however are not directly obtainable from this computational procedure.

Alternatively θ may be estimated directly from (2.22) under (2.23). The conditional distribution (2.4) contains further information on the covariance parameters γ above that provided by the marginal distribution of y_2. Let $Q_i = [Q_{i1}{}^{p\times q_i}\ Q_{i2}{}^{p\times(p-q_i)}]$, $i = 1,\ldots,m$, be $p \times p$ matrices of full column rank such that

$$Q_{i1}'M_i = I_{q_i},\quad Q_{i2}'M_i = 0,\quad i = 1,\ldots,m$$

holds. Define

$$Q = \begin{bmatrix} Q_{11} & 0 & \cdots & 0 & Q_{12} & 0 & \cdots & 0 \\ 0 & Q_{21} & \cdots & 0 & 0 & Q_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Q_{m1} & 0 & 0 & \cdots & Q_{m2} \end{bmatrix} = [Q_1\ Q_2].$$

If we let

$$L_\theta' = \begin{bmatrix} Q_1' & 0 \\ Q_2' & 0 \\ 0 & I_{(N-mp)} \end{bmatrix} L' = \begin{bmatrix} Q_1'L_1' \\ Q_2'L_1' \\ L_2' \end{bmatrix} = \begin{bmatrix} L_{11}' \\ L_{12}' \\ L_2' \end{bmatrix},$$

which has full rank, and $\boldsymbol{X}^*$ is the $N \times q$ ($q = \sum_{i=1}^{m} q_i$) matrix, $\boldsymbol{XA}$, then

$$\boldsymbol{L}_{11}'\boldsymbol{X}^* = \boldsymbol{I}^{q\times q} \quad \text{and} \quad [\boldsymbol{L}_{12}\ \boldsymbol{L}_2]'\boldsymbol{X}^* = \boldsymbol{0}.$$

If $\boldsymbol{y}_{11}^* = \boldsymbol{L}_{11}'\boldsymbol{y}$, $\boldsymbol{y}_{12}^* = \boldsymbol{L}_{12}'\boldsymbol{y}$ and $\boldsymbol{y}_2^* = \boldsymbol{L}_2'\boldsymbol{y} = \boldsymbol{y}_2$ then the transformed matrix $\boldsymbol{L}_\theta'\boldsymbol{y}$ is distributed as

$$\boldsymbol{L}_\theta'\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_{11}^* \\ \boldsymbol{y}_{12}^* \\ \boldsymbol{y}_2^* \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\theta} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{L}'_{11}\boldsymbol{\Omega}\boldsymbol{L}_{11} & \boldsymbol{L}'_{11}\boldsymbol{\Omega}\boldsymbol{L}_{12} & \boldsymbol{L}'_{11}\boldsymbol{\Omega}\boldsymbol{L}_2 \\ \boldsymbol{L}'_{12}\boldsymbol{\Omega}\boldsymbol{L}_{11} & \boldsymbol{L}'_{12}\boldsymbol{\Omega}\boldsymbol{L}_{12} & \boldsymbol{L}'_{12}\boldsymbol{\Omega}\boldsymbol{L}_2 \\ \boldsymbol{L}'_2\boldsymbol{\Omega}\boldsymbol{L}_{11} & \boldsymbol{L}'_2\boldsymbol{\Omega}\boldsymbol{L}_{12} & \boldsymbol{L}'_2\boldsymbol{\Omega}\boldsymbol{L}_2 \end{bmatrix} \right).$$

The simple maximum likelihood estimate of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = ((\boldsymbol{X}^*)'\boldsymbol{\Omega}\boldsymbol{X}^*)^{-1}(\boldsymbol{X}^*)'\boldsymbol{\Omega}^{-1}\boldsymbol{y} = (\boldsymbol{A}'\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{y},$$

from the conditional distribution of $\boldsymbol{y}_{11}^*$ given $[\boldsymbol{y}_{12}^*\ \boldsymbol{y}_2^*]$. This follows by applying the results of §2.2 with $\boldsymbol{X}^*$, $\boldsymbol{L}_{11}$, $[\boldsymbol{L}_{12}\ \boldsymbol{L}_2]$, $\boldsymbol{y}_{11}^*$ and $[\boldsymbol{y}_{12}^*\ \boldsymbol{y}_2^*]$ replacing $\boldsymbol{X}$, $\boldsymbol{L}_1$, $\boldsymbol{L}_2$, $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ respectively, and N and q replacing n and p respectively. Equation (2.5) can now be used to estimate $\boldsymbol{\theta}$.

The full REML estimate of the covariance parameters is derived from the marginal distribution of $[\boldsymbol{y}_{12}^*\ \boldsymbol{y}_2^*]$. Estimating the covariance parameters from the marginal distribution of $\boldsymbol{y}_2^*$ alone leads to a partial REML estimate of the covariance parameters. In chapter 3 we will consider full and partial REML estimates for the profile model.

Chapter 3

The Profile Model with Maximum Likelihood and REML estimation

3.1 Introduction

In this chapter we compare maximum likelihood (ML) and two versions of residual maximum likelihood (REML) for the estimation of the parameters of the profile model introduced in chapter 2. Note however, that the closing remarks of chapter 2 show that the sum of profiles model also accommodates both approaches.

The canonical form (2.19) provides a natural way to select REML components for the variance parameters. We are specifically interested in the estimation of Θ by the three likelihood methods and the estimated covariance matrix of the respective estimators of Θ. The three likelihood methods considered are,

(a) full ML, using all components of the canonical form,

(b) partial REML (PREML) using $L_2'YQ = [Y_{21}\ Y_{22}]$ for the estimation of the variance parameters and

(c) full REML using $[Y_{21}\ Y_{22}]$ and Y_{12} for the estimation of the variance parameters.

The components $[\boldsymbol{Y}_{21}\ \boldsymbol{Y}_{22}]$ and $\boldsymbol{Y}_{12}$ provide no information on $\boldsymbol{\Theta}$, have zero means, and consequently correspond to error contrasts. Consequently the variance parameters are estimated from these components.

There are two versions of REML; there is a justification for the second, PREML. Usually the profile model specified by $\boldsymbol{M}$ is unknown but the design $\boldsymbol{X}$ is determined by the experimental protocol. If we construct the marginal likelihood free of fixed effects for the regression model of $\boldsymbol{Y}$ on $\boldsymbol{X}$, we obtain PREML. This likelihood is not affected by incorrect specification of the profile model.

3.2 Estimation when the covariance matrix is unknown

For all three methods, the conditional distribution of $\boldsymbol{Y}_{11}$ given $\boldsymbol{Y}_{12}$ is used to estimate $\boldsymbol{\Theta}$; the estimator is

$$\hat{\boldsymbol{\Theta}} = \boldsymbol{Y}_{11} - \boldsymbol{Y}_{12}\hat{\boldsymbol{B}} \tag{3.1}$$

where $\hat{\boldsymbol{B}}$ is an estimator of $\boldsymbol{B}$. Under ML, estimation of $\boldsymbol{B}$ is based on the two conditional distributions of $\boldsymbol{Y}_{11}$ given $\boldsymbol{Y}_{12}$ and $\boldsymbol{Y}_{21}$ given $\boldsymbol{Y}_{22}$. The normal equations are

$$\boldsymbol{A}'\boldsymbol{W}^{-1}\boldsymbol{A}\boldsymbol{\Xi} = \boldsymbol{A}'\boldsymbol{W}^{-1}\boldsymbol{Y}_{\cdot 1},$$

where

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I}_m & \boldsymbol{Y}_{12} \\ \boldsymbol{0} & \boldsymbol{Y}_{22} \end{bmatrix}, \quad \boldsymbol{W} = \begin{bmatrix} (\boldsymbol{X}'\boldsymbol{X})^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{n-m} \end{bmatrix}, \quad \boldsymbol{\Xi} = \begin{bmatrix} \boldsymbol{\Theta} \\ \boldsymbol{B} \end{bmatrix} \quad \text{and} \quad \boldsymbol{Y}_{\cdot 1} = \begin{bmatrix} \boldsymbol{Y}_{11} \\ \boldsymbol{Y}_{21} \end{bmatrix}.$$

Hence

$$\hat{\boldsymbol{B}} = (\boldsymbol{Y}_{22}'\boldsymbol{Y}_{22})^{-1}\boldsymbol{Y}_{22}'\boldsymbol{Y}_{21} \tag{3.2}$$

and if $\boldsymbol{S} = \boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P}_X)\boldsymbol{Y}$, after manipulation (3.2) reduces to

$$\hat{\boldsymbol{B}} = (\boldsymbol{Q}_2'\boldsymbol{S}\boldsymbol{Q}_2)^{-1}\boldsymbol{Q}_2'\boldsymbol{S}\boldsymbol{Q}_1. \tag{3.3}$$

PREML and REML estimators of $\boldsymbol{B}$ are found using the conditional distribution $\boldsymbol{Y}_{21}$ given $\boldsymbol{Y}_{22}$ and hence we also obtain (3.2) for these methods. Substituting (3.3) in (3.1) leads to

$$\hat{\boldsymbol{\Theta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{S}^{-1}\boldsymbol{M}(\boldsymbol{M}'\boldsymbol{S}^{-1}\boldsymbol{M})^{-1},$$

which is the ML estimator derived by Khatri (1966)[63]. All three methods lead to the same estimators of $\boldsymbol{\Theta}$ and $\boldsymbol{B}$.

As estimation of $\boldsymbol{\Theta}$ and $\boldsymbol{B}$ involves the conditional distributions of $\boldsymbol{Y}_{11}$ given $\boldsymbol{Y}_{12}$, and $\boldsymbol{Y}_{21}$ given $\boldsymbol{Y}_{22}$, the appropriate covariance matrix of $\hat{\boldsymbol{\Theta}}$ for all three methods is the conditional covariance matrix (see also Grizzle and Allen, (1969) [51])

$$\text{var}(\hat{\boldsymbol{\Theta}}|\boldsymbol{Y}_{12}, \boldsymbol{Y}_{22}) = \boldsymbol{R} \otimes (\boldsymbol{M}'\boldsymbol{\Omega}^{-1}\boldsymbol{M})^{-1} \tag{3.4}$$

where

$$\boldsymbol{R} = (\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}'\boldsymbol{Y}_{\cdot 2}(\boldsymbol{Y}_{\cdot 2}'\boldsymbol{Y}_{\cdot 2})^{-1}\boldsymbol{Y}_{\cdot 2}'\boldsymbol{X})^{-1} \quad \text{and} \quad \boldsymbol{Y}_{\cdot 2} = \boldsymbol{Y}\boldsymbol{Q}_2.$$

We need to estimate $(\boldsymbol{M}'\boldsymbol{\Omega}^{-1}\boldsymbol{M})'$ in order to estimate (3.4). The three methods differ in the estimates of $(\boldsymbol{M}'\boldsymbol{\Omega}^{-1}\boldsymbol{M})'$. Under ML, the estimator for $\boldsymbol{\Omega}$ is $\hat{\boldsymbol{\Omega}} = \boldsymbol{S}_0/n$ where $\boldsymbol{S}_0 = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\Theta}}\boldsymbol{M}')'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\Theta}}\boldsymbol{M}')$. It can be shown that $\boldsymbol{S}_0^{-1}\boldsymbol{M} = \boldsymbol{S}^{-1}\boldsymbol{M}$ (Verbyla (1986) [147] (p. 13)) and hence the ML estimator is

$$(\boldsymbol{M}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{M})^{-1} = (\boldsymbol{M}'\boldsymbol{S}_0^{-1}\boldsymbol{M})^{-1}/n = (\boldsymbol{M}'\boldsymbol{S}^{-1}\boldsymbol{M})^{-1}/n.$$

PREML uses only $\boldsymbol{L}_2'\boldsymbol{Y}\boldsymbol{Q}$ or equivalently $\boldsymbol{L}_2'\boldsymbol{Y}$ for estimation of the variance parameters. As $\boldsymbol{L}_2'\boldsymbol{Y} \sim N(\boldsymbol{0},\ \boldsymbol{I}_{n-m} \otimes \boldsymbol{\Omega})$ the PREML estimator of $\boldsymbol{\Omega}$ is $\boldsymbol{S}/(n-m)$ which is unbiased for $\boldsymbol{\Omega}$. The PREML estimator of $\boldsymbol{M}'\boldsymbol{\Omega}^{-1}\boldsymbol{M}$ is therefore

$$(\boldsymbol{M}'\boldsymbol{S}^{-1}\boldsymbol{M})^{-1}/(n-m).$$

To estimate $(\boldsymbol{M}'\boldsymbol{\Omega}^{-1}\boldsymbol{M})^{-1}$ under REML, we use the conditional distribution of $\boldsymbol{Y}_{21}$ given $\boldsymbol{Y}_{22}$; this distribution has non-zero mean and a second application of REML is required. As in (2.18), let $\boldsymbol{T} = [\boldsymbol{T}_1^{(n-m)\times(p-q)}\ \boldsymbol{T}_2^{(n-m)\times(n-m-(p-q))}]$ have full column rank, assuming $(n-m) > (p-q)$, such that

$$\boldsymbol{T}_1'\boldsymbol{Y}_{22} = \boldsymbol{I}_{p-q}, \qquad \text{and} \qquad \boldsymbol{T}_2'\boldsymbol{Y}_{22} = \boldsymbol{0},$$

where the columns of T_2 are constructed to be orthonormal. Let $D_1 = T_1'Y_{21}$ and $D_2 = T_2'Y_{21}$. Then

$$\begin{bmatrix} D_1 \\ D_2 \end{bmatrix} | Y_{22} \sim N\left(\begin{bmatrix} B \\ 0 \end{bmatrix}, \begin{bmatrix} (Y_{22}'Y_{22})^{-1} & 0 \\ 0 & I_{n-m-(p-q)} \end{bmatrix} \otimes (M'\Omega^{-1}M)^{-1} \right),$$

and D_1 and D_2 are conditionally independent. The REML estimator of $(M'\Omega^{-1}M)^{-1}$ is therefore derived from the conditional distribution of D_2 given Y_{22}, and is

$$D_2'D_2/(n-m-(p-q)) = (M'S^{-1}M)^{-1}/(n-m-(p-q)),$$

which is unbiased for $(M'\Omega^{-1}M)^{-1}$ and is the form stated by Grizzle and Allen (1969) [51]. Notice also

$$\hat{B} = D_1 = T_1'Y_{21} = (Y_{22}'Y_{22})^{-1}Y_{22}'Y_{21}$$

as in (3.2).

All three estimators of $M'\Omega^{-1}M$ differ only in their divisors, and PREML and REML provide simple degrees of freedom adjustments to the ML estimator. PREML and REML reduce the bias of maximum likelihood estimation of the variance parameters, with REML providing the full adjustment.

3.3 Structured Covariance Matrices

3.3.1 Rao's Covariance Form

Rao (1967) [115] considered a covariance matrix of the form

$$\Omega = M\Sigma_1M' + Z\Sigma_2Z' + \sigma^2I,$$

where Σ_1 and Σ_2 are possibly unknown, $M'Z = 0$. Z need not be a basis of $R(M)^{\perp}$. We consider a simpler form

$$\Omega = M\Sigma_1M' + Z\Sigma_2Z', \tag{3.5}$$

where $\boldsymbol{Z}$ is a $p \times (p-q)$ matrix of full column rank. Let $\boldsymbol{Q}_1$, $\boldsymbol{Q}_2$, $\boldsymbol{L}_1$ and $\boldsymbol{L}_2$ be matrices of full column rank that satisfy (2.18) and additionally

$$\boldsymbol{Z}'\boldsymbol{Q}_1 = \boldsymbol{0}, \qquad \text{and} \qquad \boldsymbol{Z}'\boldsymbol{Q}_2 = \boldsymbol{I}_{p-q}.$$

These restrictions on $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ mean $\boldsymbol{Q}_1 = \boldsymbol{M}(\boldsymbol{M}'\boldsymbol{M})^{-1}$ and $\boldsymbol{Q}_2 = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$. Hence the canonical form reduces to

$$\boldsymbol{L}'\boldsymbol{Y}\boldsymbol{Q} \sim N\left(\begin{bmatrix} \boldsymbol{\Theta} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} (\boldsymbol{X}'\boldsymbol{X})^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{n-m} \end{bmatrix} \otimes \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_2 \end{bmatrix}\right),$$

so that $\boldsymbol{Y}_{11}, \boldsymbol{Y}_{12}, \boldsymbol{Y}_{21}$ and $\boldsymbol{Y}_{22}$ are independent and inferences concerning $\boldsymbol{\Theta}$ are based on the marginal distribution of $\boldsymbol{Y}_{11}$. The ML estimate of $\boldsymbol{\Theta}$ is then

$$\hat{\boldsymbol{\Theta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{M}(\boldsymbol{M}'\boldsymbol{M})^{-1},$$

which does not depend upon $\boldsymbol{\Omega}$. Thus the ML estimator is also the PREML and REML estimator.

The ML estimators of $\boldsymbol{\Sigma}_1$ (via $\boldsymbol{L}'\boldsymbol{Y}\boldsymbol{Q}_1$ or equivalently $\boldsymbol{Y}\boldsymbol{Q}_1$) and $\boldsymbol{\Sigma}_2$ (via $\boldsymbol{Y}\boldsymbol{Q}_2$) are

$$\hat{\boldsymbol{\Sigma}}_1 = (\boldsymbol{M}'\boldsymbol{M})^{-1}\boldsymbol{M}'\boldsymbol{S}\boldsymbol{M}(\boldsymbol{M}'\boldsymbol{M})^{-1}/n, \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}_2 = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}'\boldsymbol{Y}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}/n.$$

The PREML estimators of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are the ML estimators with n replaced by $n-m$. The REML estimators have divisors $n-m$ and n for $\hat{\boldsymbol{\Sigma}}_1$ and $\hat{\boldsymbol{\Sigma}}_2$ respectively. The PREML and REML estimators of $\boldsymbol{\Sigma}_1$ are unbiased and the ML and REML (but not PREML) estimators for $\boldsymbol{\Sigma}_2$ are unbiased. The covariance matrix of $\hat{\boldsymbol{\Theta}}$ is $(\boldsymbol{X}'\boldsymbol{X})^{-1} \otimes (\boldsymbol{M}'\boldsymbol{M})^{-1}\boldsymbol{M}'\boldsymbol{\Omega}\boldsymbol{M}(\boldsymbol{M}'\boldsymbol{M})^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1} \otimes \boldsymbol{\Sigma}_1$ for all three methods. PREML and REML again provide the degrees of freedom corrections for estimation of $\boldsymbol{\Sigma}_1$.

The modified model,

$$\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\Theta}_0\boldsymbol{M}_0', \ \boldsymbol{I}_n \otimes \boldsymbol{\Omega}),$$

where $\boldsymbol{M}_0$ is a known $p \times q_0$ matrix of full column rank, $q_0 < q$, and $\boldsymbol{\Omega}$ is (3.5) as before, may arise when individual sampling units are represented by, for example, higher degree polynomials than that which adequately describes the overall mean; see Fearn (1977)

[40]. Let $\boldsymbol{M}_1$ be a $p \times q_1$ matrix such that $q_0 + q_1 = q$ and $\boldsymbol{M} = [\boldsymbol{M}_0\ \boldsymbol{M}_1]$ be of full column rank. Since $\boldsymbol{YQ}_1$ and $\boldsymbol{YQ}_2$ are independent, inferences on $\boldsymbol{\Theta}_0$ are based on the marginal distribution

$$\boldsymbol{YQ}_1 \sim N([\boldsymbol{\Theta}_0\ \boldsymbol{0}],\ \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}_1).$$

If $\boldsymbol{M}^* = (\boldsymbol{M}'\boldsymbol{M})^{-1}\boldsymbol{M}'\boldsymbol{M}_0$, a $q \times q_0$ matrix of full column rank, which incidentally is $[\boldsymbol{I}_{q_0}\ \boldsymbol{0}']'$, then

$$\boldsymbol{YQ}_1 \sim N(\boldsymbol{X\Theta}_0(\boldsymbol{M}^*)',\ \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}_1)$$

which has a profile model form and the results of §2.5 can be applied for estimating $\boldsymbol{\Theta}_0$.

3.3.2 The General Case

Suppose $\boldsymbol{\Omega} = \sigma^2\boldsymbol{V}(\gamma)$, where γ is an $s \times 1$ vector of unknown parameters with jth component γ_j. This section is primarily concerned with estimation of σ^2 and γ. All three approaches give (2.20) as the estimator of $\hat{\boldsymbol{\Theta}}$ with γ replaced by the appropriate estimate. The likelihood functions for estimation of σ^2 and γ for the three approaches are found using the canonical form (2.19), although for maximum likelihood it is more convenient to use the original growth curve model. The three log-likelihoods for estimation of σ^2 and γ are

$$\begin{aligned}
L_M &= -\frac{np}{2}\log\sigma^2 - \frac{n}{2}\log(\det \boldsymbol{V}) - \frac{1}{2\sigma^2}\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{R}_0),\\
L_{PR} &= -\frac{(n-m)p}{2}\log\sigma^2 - \frac{n-m}{2}\log(\det \boldsymbol{V}) - \frac{1}{2\sigma^2}\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{S}),\\
L_R &= -\frac{np-mq}{2}\log\sigma^2 - \frac{n}{2}\log(\det \boldsymbol{V}) - \frac{m}{2}\det(\boldsymbol{M}'\boldsymbol{V}^{-1}\boldsymbol{M}) - \frac{1}{2\sigma^2}\mathrm{tr}(\boldsymbol{V}^{-1}\hat{\boldsymbol{R}}_0),
\end{aligned}$$

where $\boldsymbol{R}_0 = (\boldsymbol{Y} - \boldsymbol{X\Theta M}')(\boldsymbol{Y} - \boldsymbol{X\Theta M}')'$, $\hat{\boldsymbol{R}}_0$ is $\boldsymbol{R}_0$ with $\boldsymbol{\Theta}$ replaced by (2.20) which is a function of γ, and $\boldsymbol{S}$ is given preceding (3.3). The estimates of σ^2 are all of the form

$$\hat{\sigma}^2 = \frac{1}{\nu}\mathrm{tr}(\hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{R}}_0) \tag{3.6}$$

where ν is np, $(n-m)p$ and $np-mq$ for ML, PREML and REML respectively, $\hat{\boldsymbol{V}} = \boldsymbol{V}(\hat{\gamma})$ is evaluated at the ML, PREML or REML estimate of γ and $\hat{\boldsymbol{R}}_0$ is $\boldsymbol{S}_0$ evaluated at $\hat{\boldsymbol{\Theta}}$

for ML and REML, and $\boldsymbol{S}$ for PREML. Note that we are using $\hat{\boldsymbol{R}}_o$ instead of $\boldsymbol{R}_o$ in L_R because using REML, the mean parameter matrix $\boldsymbol{\Theta}$ is replaced in the error contrast likelihood function by the estimate $\hat{\boldsymbol{\Theta}}(\boldsymbol{\gamma})$. Further, the expectation of the score equations is zero in L_R by noting that

$$E(\hat{\boldsymbol{R}}_0) = \sigma^2 n\boldsymbol{V} - \sigma^2 m\boldsymbol{M}'(\boldsymbol{M}'\boldsymbol{V}^{-1}\boldsymbol{M})^{-1}\boldsymbol{M}'.$$

This is not the case if $\boldsymbol{R}_o$ was used instead in L_R. The REML log-likelihood L_R has adjustment

$$-\frac{m}{2}\det(\boldsymbol{M}'\boldsymbol{V}^{-1}\boldsymbol{M})$$

to the profile likelihood under ML.

Let $\boldsymbol{V}^{(j)} = \partial\boldsymbol{V}^{-1}/\partial\gamma_j$. Under ML the score vector has jth component, $j = 1, \ldots, s$,

$$\frac{\partial L_M}{\partial\gamma_j} = \frac{n}{2}\mathrm{tr}(\boldsymbol{V}^{(j)}\boldsymbol{V}) - \frac{1}{2\sigma^2}\mathrm{tr}(\boldsymbol{V}^{(j)}\boldsymbol{S}_o).$$

For PREML, the only change involves the replacement of n by $n-m$ and $\boldsymbol{S}_o$ by $\boldsymbol{S}$. For REML, there is an additional term to those in the ML result namely

$$-\frac{m}{2}\mathrm{tr}(\boldsymbol{M}'\boldsymbol{V}^{(j)}\boldsymbol{M}(\boldsymbol{M}'\boldsymbol{V}^{-1}\boldsymbol{M})^{-1}) = -\frac{m}{2}\mathrm{tr}(\boldsymbol{M}^{(j)}), \text{ say.}$$

The expected information matrix from the full likelihood has components

$$\begin{array}{c} \\ \sigma^2 \\ \gamma_i \\ \gamma_j \end{array}\begin{array}{c} \begin{array}{ccc} \sigma^2 & \gamma_i & \gamma_j \end{array} \\ \begin{pmatrix} \frac{np}{2\sigma^4} & \frac{n}{2\sigma^2}\mathrm{tr}(\boldsymbol{V}^{(i)}\boldsymbol{V}) & \frac{n}{2\sigma^2}\mathrm{tr}(\boldsymbol{V}^{(j)}\boldsymbol{V}) \\ & \frac{n}{2}\mathrm{tr}(\boldsymbol{V}^{(i)}\boldsymbol{V}\boldsymbol{V}^{(i)}\boldsymbol{V}) & \frac{n}{2}\mathrm{tr}(\boldsymbol{V}^{(i)}\boldsymbol{V}\boldsymbol{V}^{(j)}\boldsymbol{V}) \\ & & \frac{n}{2}\mathrm{tr}(\boldsymbol{V}^{(j)}\boldsymbol{V}\boldsymbol{V}^{(j)}\boldsymbol{V}) \end{pmatrix} \end{array}.$$

Under PREML and REML, np in the expected information for σ^2 is replaced by $(n-m)p$ and $np-mq$ respectively, while n in the remaining terms is replaced by $n-m$ under PREML. Under REML additional terms like

$$\begin{array}{ccc} (\sigma^2,\ \gamma_i) & (\gamma_i,\ \gamma_i) & (\gamma_i,\ \gamma_j) \\ -\frac{m}{2\sigma^2}\mathrm{tr}(\boldsymbol{M}^{(i)}), & -\frac{m}{2}\mathrm{tr}(\boldsymbol{M}^{(i)}\boldsymbol{M}^{(i)}), & -\frac{m}{2}\mathrm{tr}(\boldsymbol{M}^{(i)}\boldsymbol{M}^{(j)}) \end{array}$$

are required.

We can use the method of scoring to estimate γ for all three methods; in fact as we have an explicit expression for $\hat{\sigma}^2$ in each case, we need only consider the scoring equation (2.15).

The asymptotic covariance matrix of all three estimators of Θ is given by (2.21). No analytical solutions appear to exist for the conditional covariance matrix $\text{var}(\hat{\Theta}|\boldsymbol{Y}_{12})$ because $\hat{\Theta}$ is a complex nonlinear function of $\hat{\gamma}$. Therefore, to make a comparison of the three estimators, we use the asymptotic covariance matrix in the simulation study discussed in the following section.

3.4 A Simulation Study: First Order Autoregressive Covariance Structures

Consider the case where the covariance matrix has a first order autoregressive structure (AR-1), that is

$$\boldsymbol{\Omega} = \sigma^2 \boldsymbol{V},$$

where

$$(1-\rho^2)\boldsymbol{V} = \begin{bmatrix} 1 & \rho & \cdots & \rho^{p-1} \\ \rho & 1 & \cdots & \rho^{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \cdots & 1 \end{bmatrix}.$$

The inverse of $\boldsymbol{V}$ is $\boldsymbol{D} = (\boldsymbol{I}_p + \rho^2 \boldsymbol{E}_1 - \rho \boldsymbol{F}_1)$ (see Verbyla (1985) [145]). Here $\boldsymbol{E}_1$ is a diagonal matrix with the first and last diagonal entries set to zero and the remaining entries set to one; $\boldsymbol{F}_1$ has ones along the first upper and lower minor diagonals and zero elsewhere. As the inverse of $\boldsymbol{V}$ is a simpler expression to use than $\boldsymbol{V}$, we will work with the likelihood function written in terms of $\boldsymbol{D}$.

We already know the form of the estimator of Θ for ML when $\boldsymbol{\Omega}$ is known, that is (2.20). When $\boldsymbol{\Omega}$ is not known, estimates of the scalar variance parameters σ^2 and ρ are

required using likelihood methods (a), (b) and (c). We then substitute the respective estimators of the variance parameters into (2.20) to give estimators for $\boldsymbol{\Theta}$ for all three methods.

In the following, the subscripts M, PR and R refer to estimators under ML, PREML and REML respectively. The ML estimator for σ^2 is (3.6) ($\nu = np$), and the ML equation for ρ can be simplified to the cubic equation,

$$\frac{p}{2}\mathrm{tr}(\boldsymbol{F}_1\boldsymbol{R}_0) - (\mathrm{tr}(\boldsymbol{R}_0) + p\mathrm{tr}(\boldsymbol{E}_1\boldsymbol{R}_0))\hat{\rho}_M - \frac{p-2}{2}\mathrm{tr}(\boldsymbol{F}_1\boldsymbol{R}_0)\hat{\rho}_M^2 + (p-1)\mathrm{tr}(\boldsymbol{E}_1\boldsymbol{R}_0)\hat{\rho}_M^3 = 0\ .$$

In method (b) the PREML equation for ρ can also be represented as a cubic equation

$$\frac{p}{2}\mathrm{tr}(\boldsymbol{F}_1\boldsymbol{S}) - (\mathrm{tr}(\boldsymbol{S}) + p\mathrm{tr}(\boldsymbol{E}_1\boldsymbol{S}))\hat{\rho}_{PR} - \frac{p-2}{2}\mathrm{tr}(\boldsymbol{F}_1\boldsymbol{S})\hat{\rho}_{PR}^2 + (p-1)\mathrm{tr}(\boldsymbol{E}_1\boldsymbol{S})\hat{\rho}_{PR}^3 = 0\ .$$

The divisor for $\hat{\sigma}^2_{PR}$ is $\nu = (n-m)p$.

Finally in method (c) the likelihood equation for ρ can be written as a quintic

$$\alpha_0 - \alpha_1\rho_R - \alpha_2\rho_R^2 + \alpha_3\rho_R^3 - \alpha_4\rho_R^4 + \alpha_5\rho_R^5 = 0$$

where

$$\alpha_0 = \frac{m}{2}C\mathrm{tr}(\hat{\boldsymbol{R}}_0) + \frac{np-mq}{2}\mathrm{tr}(\boldsymbol{F}_1\hat{\boldsymbol{R}}_0)$$

$$\alpha_1 = (n+mA)\mathrm{tr}(\hat{\boldsymbol{R}}_0) + \frac{m}{2}C\mathrm{tr}(\boldsymbol{F}_1\hat{\boldsymbol{R}}_0) + (np-mq)\mathrm{tr}(\boldsymbol{E}_1\hat{\boldsymbol{R}}_0)$$

$$\alpha_2 = (n+mA)\mathrm{tr}(\boldsymbol{F}_1\hat{\boldsymbol{R}}_0) + \frac{m}{2}C\mathrm{tr}(\boldsymbol{E}_1\hat{\boldsymbol{R}}_0) - \frac{m}{2}C\mathrm{tr}(\hat{\boldsymbol{R}}_0) - \frac{np-mq}{2}\mathrm{tr}(\boldsymbol{F}_1\hat{\boldsymbol{R}}_0)$$

$$\alpha_3 = -(n+mA)\mathrm{tr}(\boldsymbol{E}_1\hat{\boldsymbol{R}}_0) + mA\mathrm{tr}(\hat{\boldsymbol{R}}_0) + \frac{m}{2}C\mathrm{tr}(\boldsymbol{F}_1\hat{\boldsymbol{R}}_0) + (np-mq)\mathrm{tr}(\boldsymbol{E}_1\hat{\boldsymbol{R}}_0)$$

$$\alpha_4 = mA\mathrm{tr}(\boldsymbol{F}_1\hat{\boldsymbol{R}}_0) + \frac{m}{2}C\mathrm{tr}(\boldsymbol{E}_1\hat{\boldsymbol{R}}_0) \quad \alpha_5 = mA\mathrm{tr}(\boldsymbol{E}_1\hat{\boldsymbol{R}}_0),$$

where we have set $A = \mathrm{tr}((\boldsymbol{M}'\boldsymbol{D}\boldsymbol{M})^{-1}\boldsymbol{M}'\boldsymbol{E}_1\boldsymbol{M})$ and $C = \mathrm{tr}((\boldsymbol{M}'\boldsymbol{D}\boldsymbol{M})^{-1}\boldsymbol{M}'\boldsymbol{F}_1\boldsymbol{M})$. The divisor for $\hat{\sigma}^2_R$ is $\nu = np - mq$.

A simple simulation study was performed so that comparisons could be made, for the autoregressive case, on the three likelihood methods considered in this chapter in terms of estimation of $\boldsymbol{\Theta}$ and the pertinent standard errors. Different values of the autoregressive parameter ρ and 'group size' m were considered. The mean was defined

to be a straight line, that is $\theta_{1j} + \theta_{2j}t$ where j is the treatment number. For simplicity, we let the parameters of the lines be identical over the different treatments, that is

$$\theta_{1j} = \theta_1 \quad \text{and} \quad \theta_{2j} = \theta_2 \quad \forall j.$$

The number of sampling units and time points are $n = 40$ and $p = 6$ respectively and we have let θ_1 and θ_2 equal 2 and 1 respectively. At each combination of values of ρ and m considered, a total of 500 simulations were performed. The number of groups, m, considered are 2, 5, 10 and 20.

We use the asymptotic covariance matrix (2.21) for all three estimators of $\boldsymbol{\Theta}$ to show that ML biases the standard errors downwards and that REML provides more suitable estimators of the standard errors. To do this we compare the estimates of the matrix $(\boldsymbol{M}'\boldsymbol{\Omega}^{-1}\boldsymbol{M})^{-1}$ for the three methods. The quadratic equation

$$\boldsymbol{x}'(\boldsymbol{M}'\boldsymbol{\Omega}^{-1}\boldsymbol{M})^{-1}\boldsymbol{x} = 1$$

defines an ellipse which can be mapped to a unit circle $\boldsymbol{y}'\boldsymbol{y} = 1$ where

$$\boldsymbol{y} = (\boldsymbol{M}'\boldsymbol{\Omega}^{-1}\boldsymbol{M})^{-\frac{1}{2}}\boldsymbol{x}.$$

Let $\hat{\rho}$ and $\hat{\sigma}^2$ be estimators of ρ and σ^2. We denote $\boldsymbol{\Omega}$ evaluated at $\hat{\rho}$ and $\hat{\sigma}^2$ by $\hat{\boldsymbol{\Omega}}$. If $\boldsymbol{x} = (\boldsymbol{M}'\boldsymbol{\Omega}^{-1}\boldsymbol{M})^{\frac{1}{2}}\boldsymbol{y}$, the quadratic equation

$$\boldsymbol{x}'(\boldsymbol{M}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{M})^{-1}\boldsymbol{x} = \boldsymbol{y}'(\boldsymbol{M}'\boldsymbol{\Omega}^{-1}\boldsymbol{M})^{\frac{1}{2}}(\boldsymbol{M}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{M})^{-1}(\boldsymbol{M}'\boldsymbol{\Omega}^{-1}\boldsymbol{M})^{\frac{1}{2}}\boldsymbol{y} = 1, \qquad (3.7)$$

defines another ellipse. If $\hat{\rho}$ and $\hat{\sigma}^2$ are unbiased, we would expect on average the ellipse (3.7) to be approximately a unit circle. If $\hat{\rho}$ and $\hat{\sigma}^2$ are biased downwards, we would expect on average the ellipse (3.7) to lie within the unit circle. The square root of the eigenvalues of $(\boldsymbol{M}'\boldsymbol{\Omega}^{-1}\boldsymbol{M})^{\frac{1}{2}}(\boldsymbol{M}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{M})^{-1}(\boldsymbol{M}'\boldsymbol{\Omega}^{-1}\boldsymbol{M})^{\frac{1}{2}}$, λ_1 and λ_2 say, give the lengths of the first and secondary principal axes of the ellipse. If the ellipse is a unit circle then both λ_1 and λ_2 are 1. When the profile model is a straight line, and we have an AR-1 covariance structure, the two eigenvectors of $(\boldsymbol{M}'\boldsymbol{\Omega}^{-1}\boldsymbol{M})^{\frac{1}{2}}(\boldsymbol{M}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{M})^{-1}(\boldsymbol{M}'\boldsymbol{\Omega}^{-1}\boldsymbol{M})^{\frac{1}{2}}$ turn out to be independent of ρ, except for the sign. With this restriction in mind we

have chosen λ_1 and λ_2 to correspond to the same eigenvectors over the 500 simulations. The lengths λ_1 and λ_2 were calculated for all 3 methods at each of the simulations, and their means were recorded over the 500 simulations. For the same set of values of ρ and m, the mean of the estimates of ρ, σ^2 and $\boldsymbol{\Theta}$ were recorded as well.

The ML, PREML and REML estimates of the mean parameters were found to differ little (see figures 3.1 and 3.2), in other words for this simple case the three likelihood estimators of $\boldsymbol{\Theta}$ considered are essentially, if not exactly, the same. It is conjectured this will be true more generally.

Figure 3.3 shows the increase of the negative bias of ML as m increases, with the bias greatest for low $|\rho|$. PREML and REML estimates are very close to the correct value of ρ. The averages of the sample standard errors of the ρ estimates are displayed in figure 3.4. The PREML standard errors rise as m increases reflecting the loss of information by not utilizing $\boldsymbol{Y}_{12}$.

The negative bias in estimating σ^2 under ML is demonstrated in figure 3.5, with the bias increasing as m increases. The bias is marginally greater for negative values of ρ. Again the PREML and REML estimates are close to the true value. The averages of the sample standard errors are recorded in figure 3.6. The standard errors under ML are low with REML rising a small amount as m gets larger. PREML standard errors rise more dramatically with increasing m, and reflects the loss of information of PREML relative to REML as was seen in figure 3.4.

Each of the diagrams in figure 3.7 represent plots of λ_1 and λ_2 against m for ML, PREML and REML. The circles correspond to the larger of the two λ's, λ_1, and the triangles correspond to λ_2. As in the previous figures PREML and REML are very close. Figure 3.7 clearly demonstrates that as m increases, ML is under-estimating the standard errors of $\hat{\boldsymbol{\Theta}}$ while PREML and REML provide more realistic standard errors. There is also an interesting trend for the ellipse defined by (3.7) to move away from a circle as ρ increases. This is occurring in a more dramatic way as m gets larger for ML. The average sample standard errors of λ_1 and λ_2 are displayed in figure 3.8. The

partial REML standard errors are slightly higher than REML again reflecting the loss of information. There is a dramatic difference in the standard error pattern for ML relative to either of the REML estimates as m increases.

As noted, PREML and REML estimators of σ^2 and ρ are very similar when the number of groups are very large. This is surprising as PREML does not use $\boldsymbol{Y}_{12}$ in its estimation of the variance parameters, which amounts to discarding information in the error space. As m increases, $\boldsymbol{Y}_{12}$ increases in size and the amount of discarded information increases.

3.5 Discussion

It appears from the cases examined in this chapter that the three likelihood methods considered all give identical or very close estimates of $\boldsymbol{\Theta}$. However ML estimation can lead to quite misleading standard errors for $\hat{\boldsymbol{\Theta}}$, where the correct covariance matrix is the conditional one. In the unstructured case, PREML and REML provide degrees of freedom corrections to $(\boldsymbol{M}'\boldsymbol{\Omega}^{-1}\boldsymbol{M})^{-1}$, REML being unbiased. PREML and more specifically REML provide more realistic levels for the standard errors of $\hat{\boldsymbol{\Theta}}$. With Rao's covariance structure we saw that PREML and REML estimators of $\boldsymbol{\Sigma}_1$ are the same and unbiased.

In the structured case, ML demonstrates an increasing negative bias for the estimation of the variance parameters as m increases, while REML provides a sounder and more robust estimating procedure. If the variance parameters are of interest, then as m increases REML provides the sensible method. Due to the reduction in bias in estimating the variance parameters, the estimates of the standard errors of $\hat{\boldsymbol{\Theta}}$ under REML will be more realistic than under ML. On this point alone, it is better to use REML than ML.

The simulations in the autoregressive case showed little difference between PREML and REML with respect to the estimates of the standard errors for the estimator of $\boldsymbol{\Theta}$. The benefits of REML as opposed to PREML require further study but the gain in

precision for the standard errors appears small. REML presumes a correct specification of the profile design matrix $\boldsymbol{M}$, and if $\boldsymbol{M}$ is incorrectly specified then an element of negative bias may occur as with ML. On the other hand PREML results in a positive bias if $\boldsymbol{M}$ is correct, but this may be a price worth paying for robustness in the face of possible incorrect specification of $\boldsymbol{M}$. A suggestion for future work is a study of the impact of $\boldsymbol{M}$ on the three likelihood methods.

Another positive point of PREML is its simplicity. As pointed out by Swallow and Monahan (1984) [140] a common problem of ML and to a greater extent REML is that variance parameter estimators need to be obtained iteratively. The resulting computing may be very complicated and time consuming. PREML can be an iterative method but the likelihood form is often much simpler and consequently easier to implement from a computing point of view.

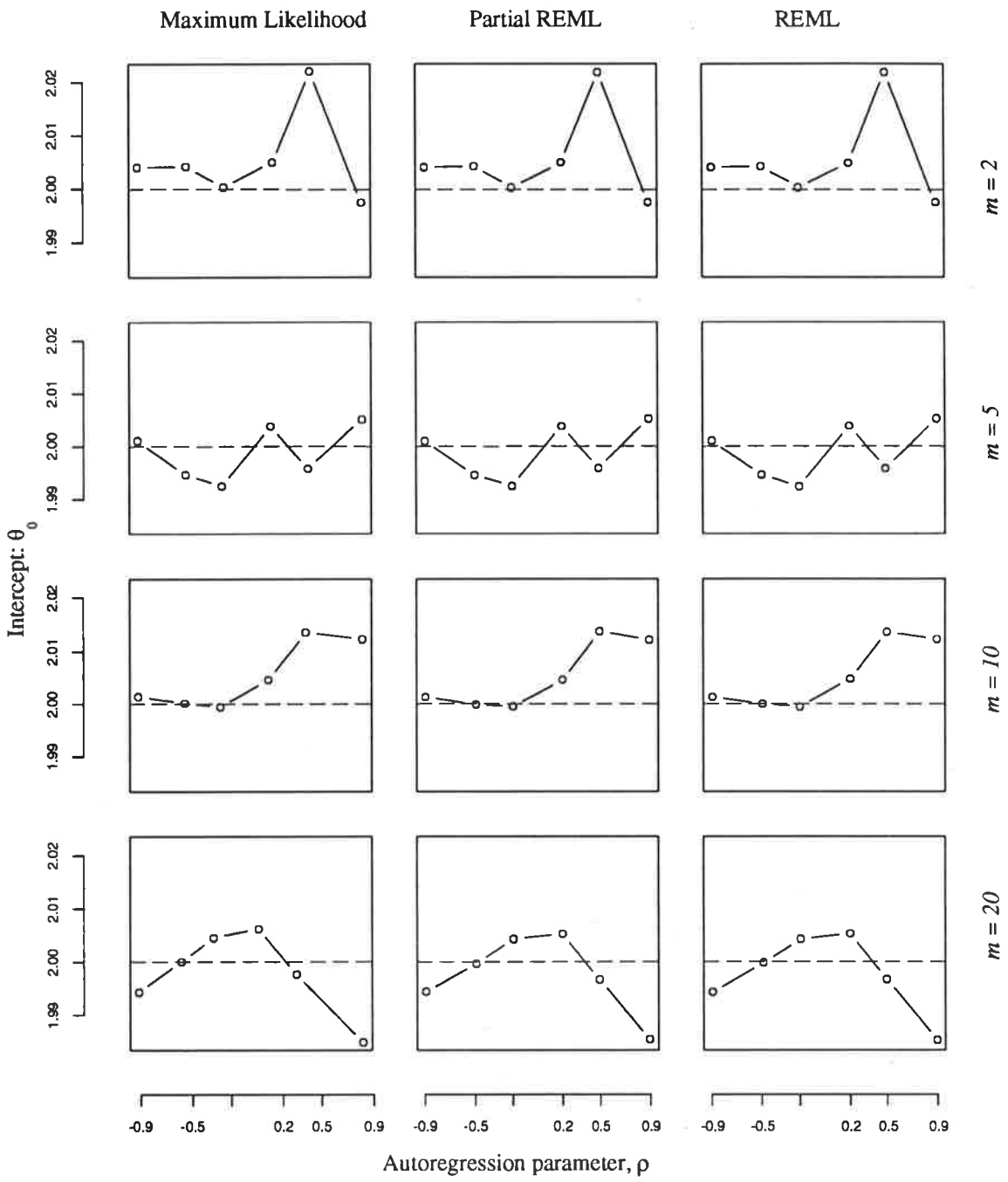

Figure 3.1: The estimate of θ_0.

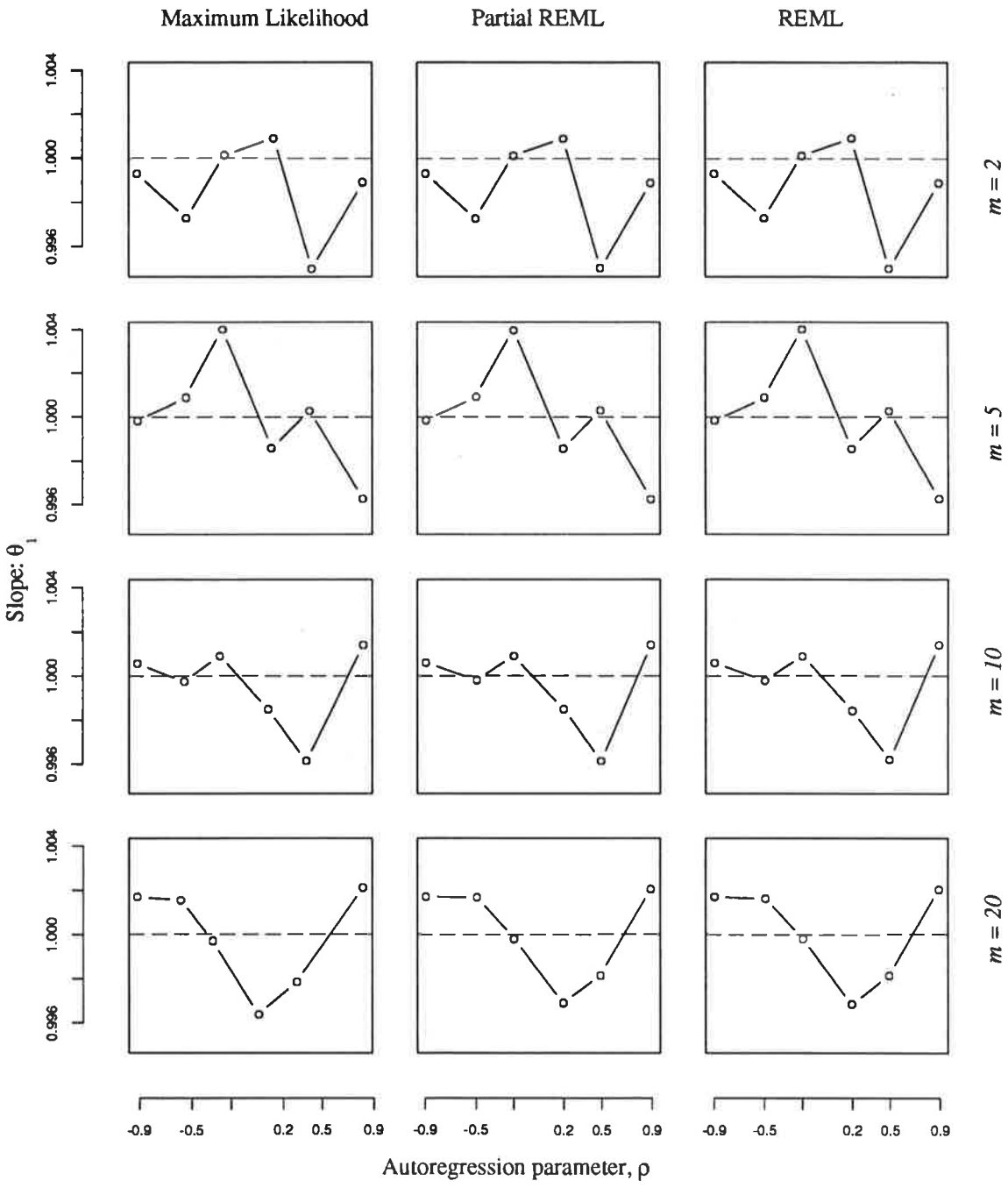

Figure 3.2: The estimate of θ_1.

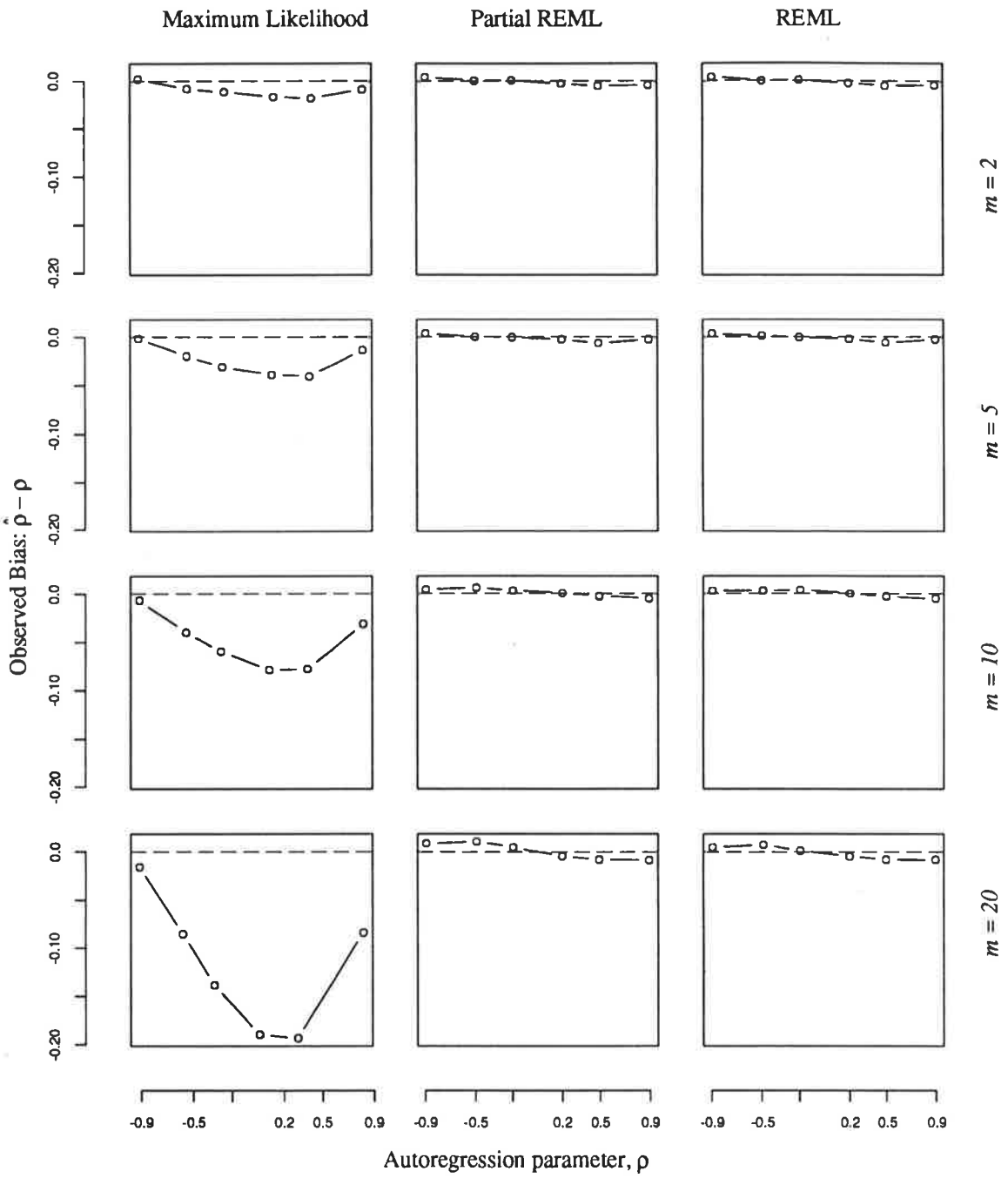

Figure 3.3: Bias in the estimate of the autoregression parameter, ρ.

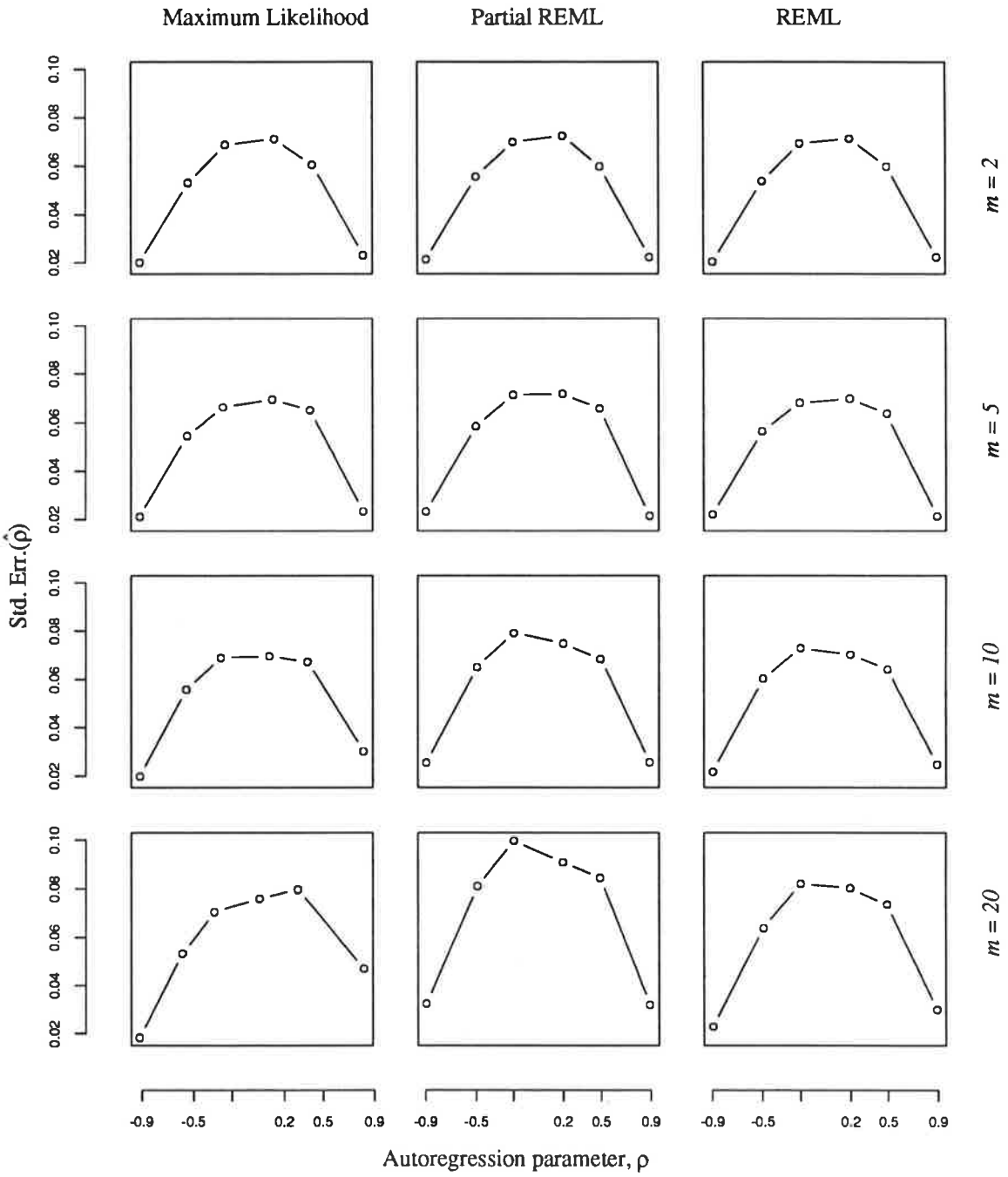

Figure 3.4: The standard errors of the estimate of ρ.

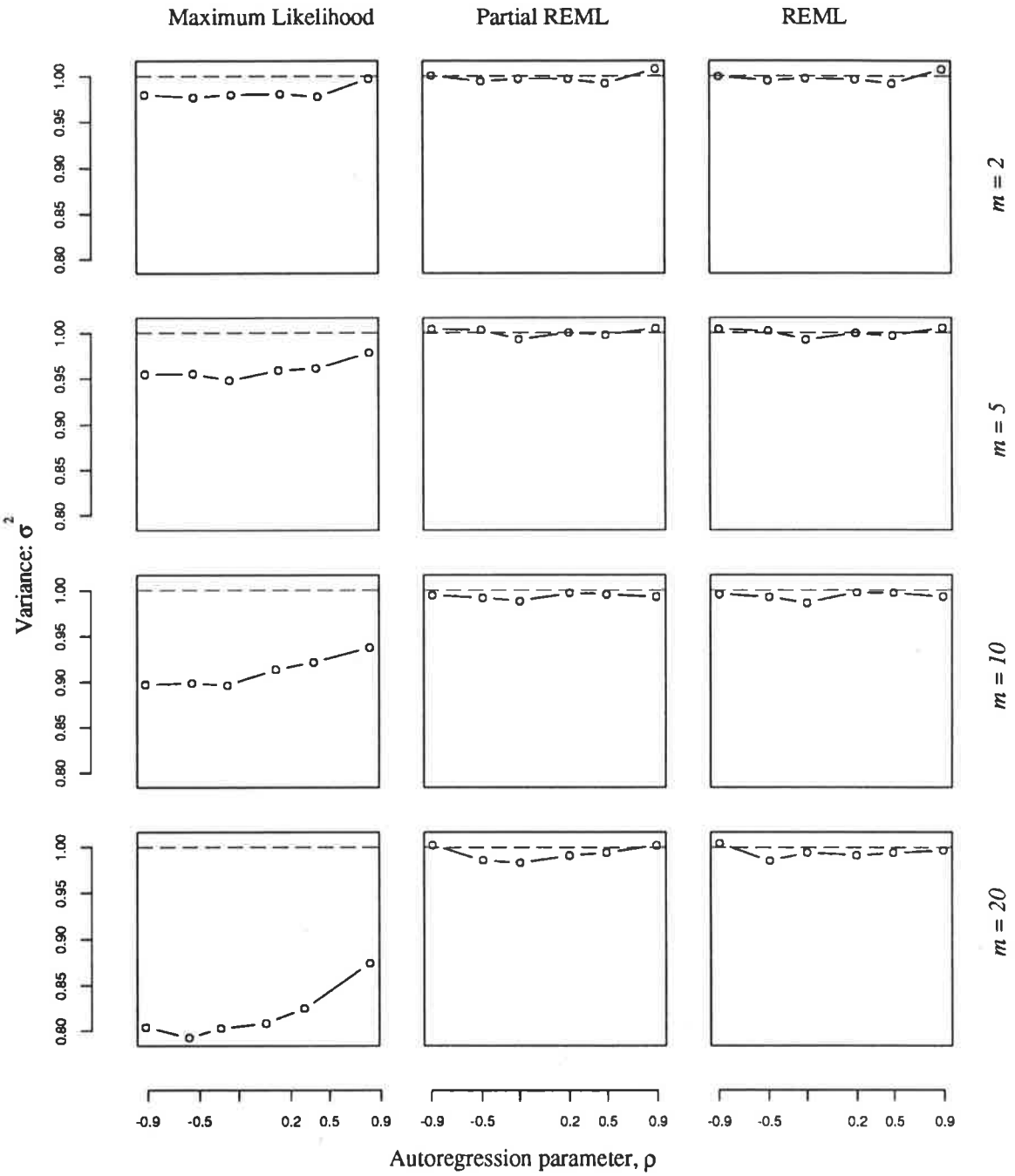

Figure 3.5: The estimate of σ^2.

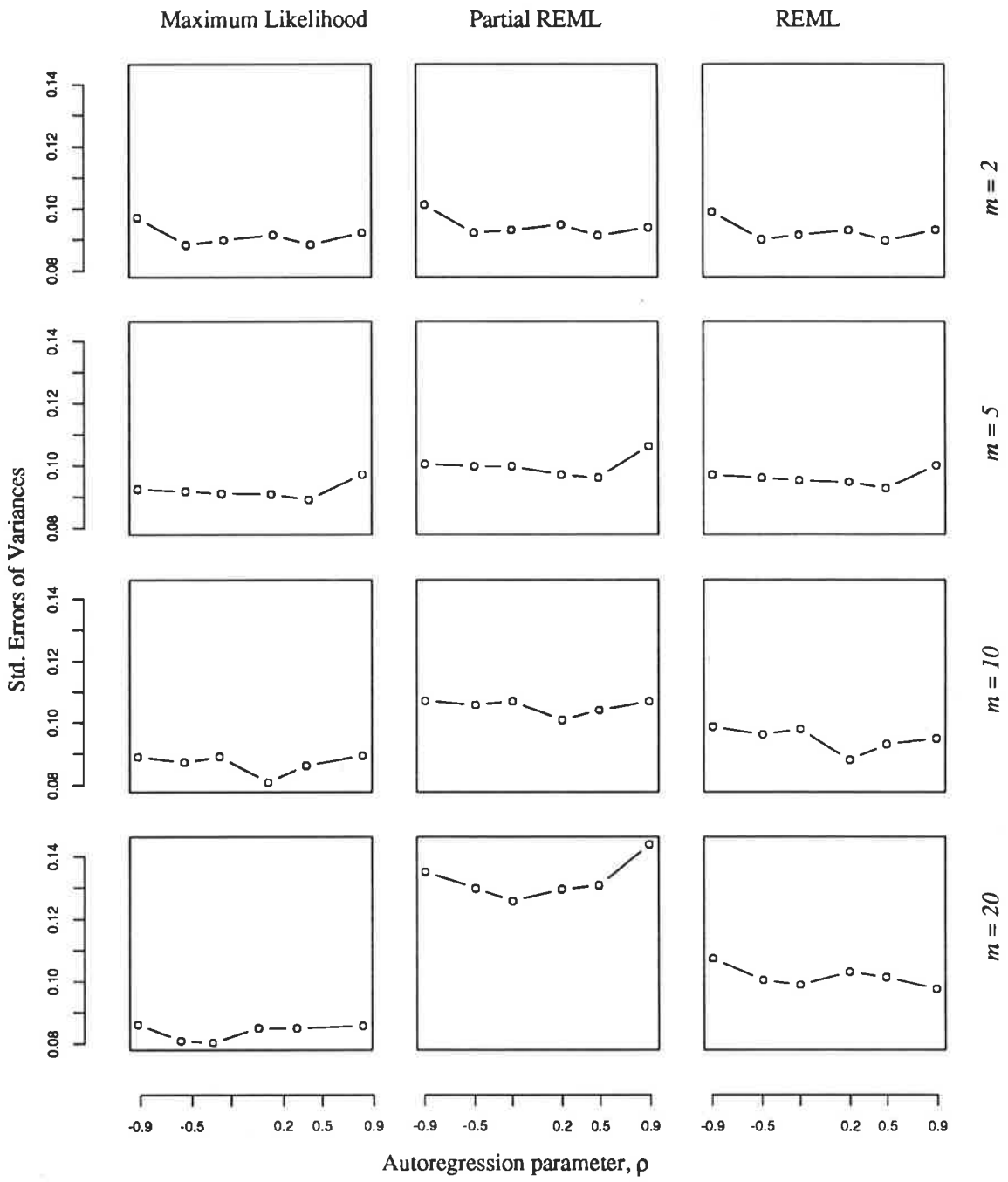

Figure 3.6: The standard errors of the estimate of σ^2.

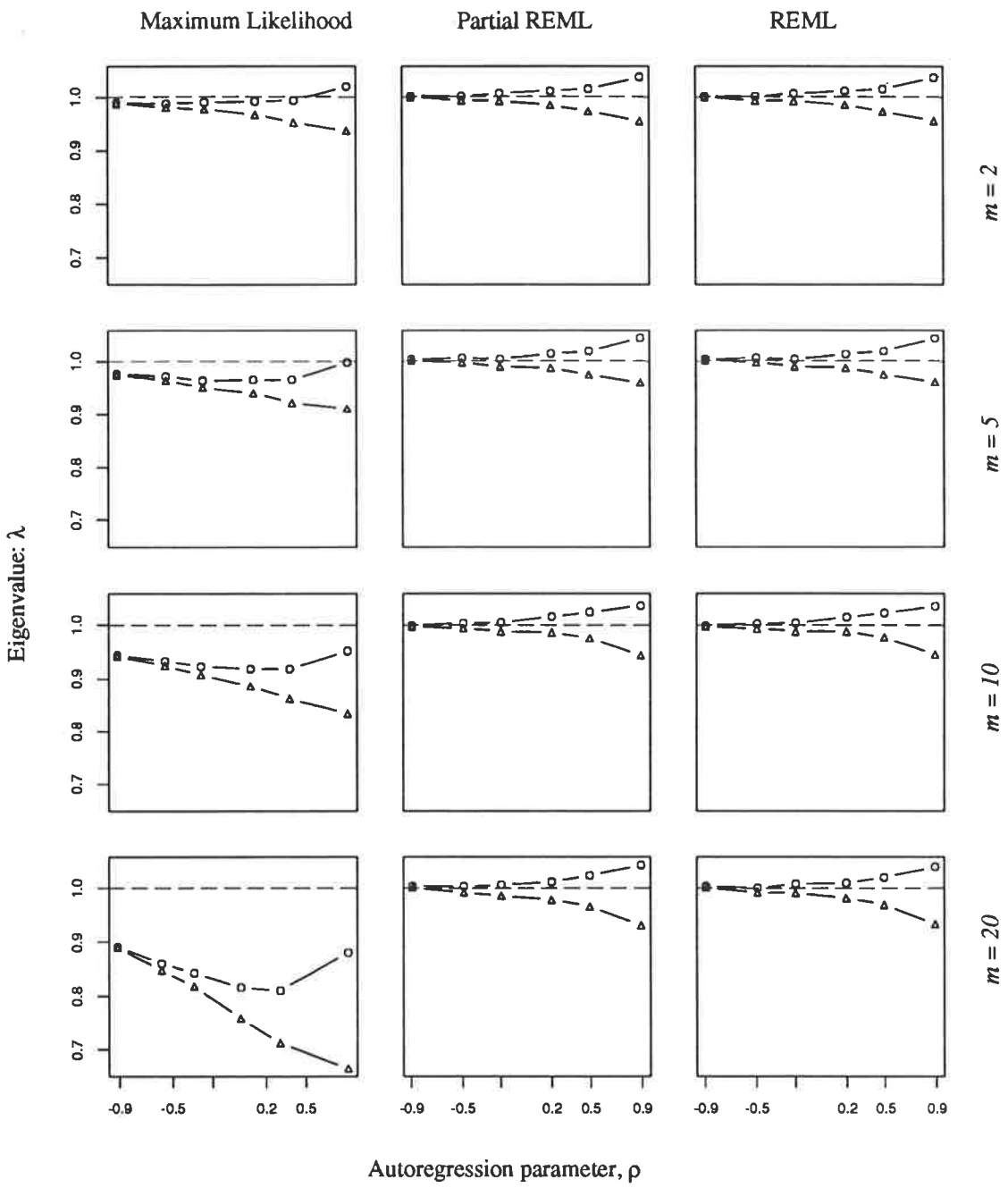

Figure 3.7: The estimates of the λ's.

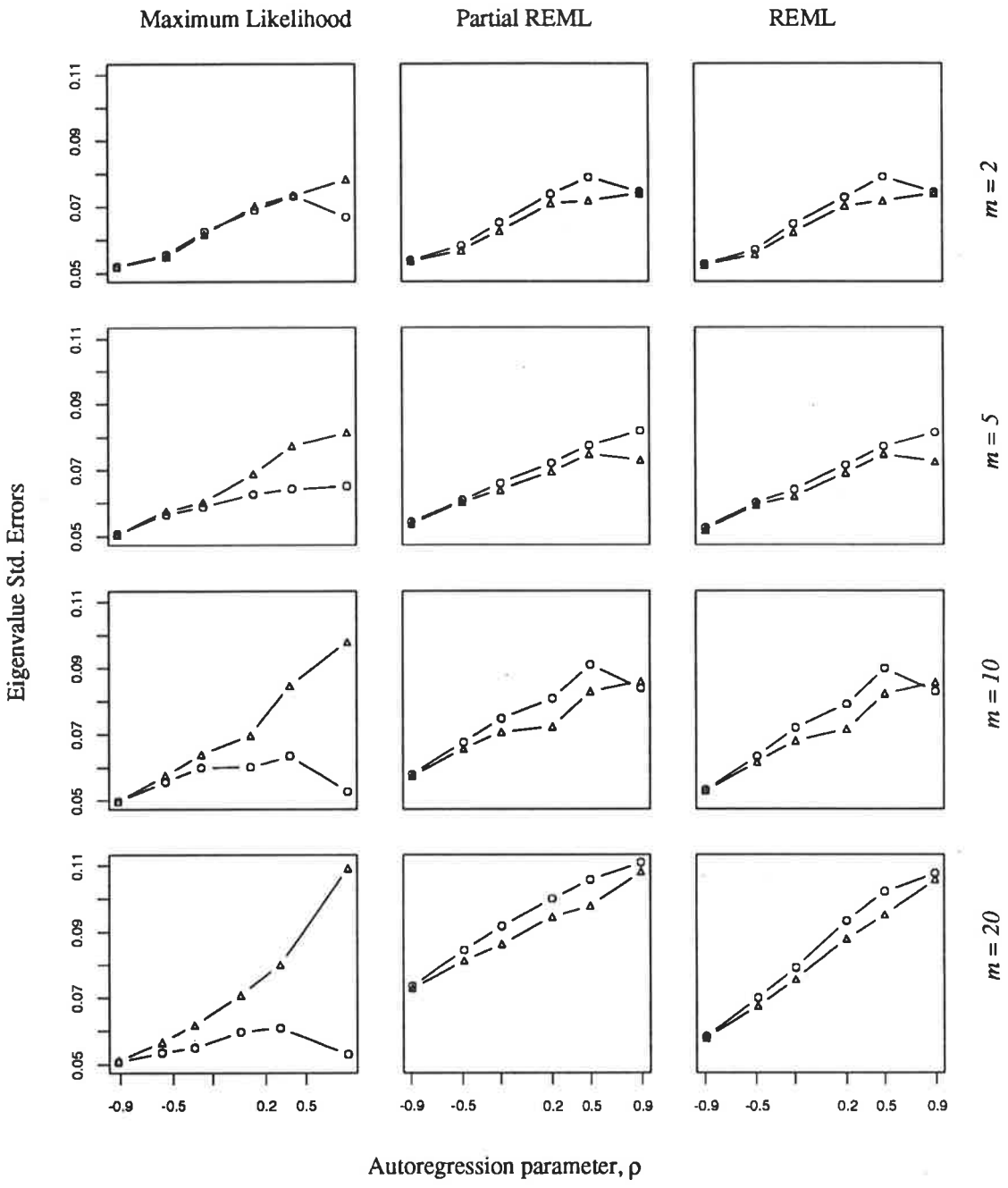

Figure 3.8: The standard errors of the estimate of the λ's.

Chapter 4

Models for Repeated Categorical Response

4.1 Introduction

Koch *et al.* (1977) [66] present a general modelling approach for repeated categorical response variables analogous to the general multivariate linear model for continuous response variables. The method is an extension of the weighted least squares (WLS) method of Grizzle *et al.* (1969) [52]. Stanek and Diehl (1988) [134] further extend the WLS method of Koch *et al.* (1977) [66] to the situation where the number of repeated measurements is too large, relative to the sample size, to allow for the modelling of all the marginal response functions. They restrict themselves to the case of repeated measurements of binary response variables. Their method is analogous to growth curve models with continuous response variables using polynomials to model the response functions.

In the incomplete data case, Koch, Imrey and Reinfurt (1972) [65] provide an approach which extends the work of Grizzle *et al.* (1969) [52]. More recently, Lipsitz, Laird and Harrington (1992) [87] propose a three stage least squares approach more along the lines of the generalized estimating equation approach of Liang and Zeger (1986) [81].

Suppose we have p repeated measurements of a categorical response of c levels for n experimental units; initially we assume there are no missing values. There are therefore $r = c^p$ possible response profiles. In the case of a binary response, $r = 2^p$. Let s denote the number of treatments or subpopulations for the n experimental units, n_i is the number of experimental units in subpopulation or treatment i, and let $\boldsymbol{Y}_i$ be the $r \times 1$ vector, $(n_{i1}, n_{i2}, \ldots, n_{ir})'$, of observed frequencies for the i^{th} subpopulation. The vector $\boldsymbol{Y}_i$ is assumed to follow a multinomial distribution with some parameter vector $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{ir})'$, where π_{ij} $(j = 1, \ldots, r)$ is the probability that a randomly selected experimental unit from the i^{th} subpopulation is observed in the j^{th} response profile.

An unbiased estimator of $\boldsymbol{\pi}_i$ is the vector of observed proportions $\boldsymbol{p}_i = \boldsymbol{Y}_i/n_i$, with covariance matrix

$$\boldsymbol{V}_i = \text{var}(\boldsymbol{p}_i) = \frac{1}{n_i}(\boldsymbol{\Lambda}_i - \boldsymbol{\pi}_i\boldsymbol{\pi}_i') \tag{4.1}$$

where $\boldsymbol{\Lambda}_i$ is a diagonal matrix with elements $\boldsymbol{\pi}_i$. The maximum likelihood estimator (MLE) of $\boldsymbol{\pi}_i$ is $\boldsymbol{p}_i$, and a consistent estimator of $\boldsymbol{V}_i$, $\hat{\boldsymbol{V}}_i$, is obtained by replacing $\boldsymbol{\pi}_i$ by $\boldsymbol{p}_i$.

In practice, marginal functions of the response proportions are of interest. Let $\boldsymbol{f}_i = [f_i(\boldsymbol{p}_i)]$ be a column vector of u_i $(u_i < r)$ differentiable functions of the vector $\boldsymbol{p}_i$. For example, $\boldsymbol{f}_i$ may be a linear transformation of the form $\boldsymbol{f}_i = \boldsymbol{A}_i\boldsymbol{p}_i$ where $\boldsymbol{A}_i$ is a matrix of known constants, forming marginal probabilities of interest. Other examples of transformations are logarithmic, $\boldsymbol{f}_i = \log(\boldsymbol{p}_i)$, and the exponential, $\boldsymbol{f}_i = \exp(\boldsymbol{p}_i)$. Grizzle *et al.* (1969) [52] and Forthofer and Koch (1973) [46] demonstrated that a large number of situations may be adequately analysed using one of these three transformations. An approximate covariance matrix of $\boldsymbol{f}_i$ is the $u_i \times u_i$ matrix

$$\boldsymbol{H}_i = \boldsymbol{P}_i\boldsymbol{V}_i\boldsymbol{P}_i', \tag{4.2}$$

where $\boldsymbol{P}_i = [\partial \boldsymbol{f}_i(\boldsymbol{\pi}_i)/\partial \boldsymbol{\pi}_i]$ is a $u_i \times r$ matrix of the first partial derivatives of the functions $\boldsymbol{f}_i$ evaluated at $\boldsymbol{p}_i$. Equation (4.2) is the asymptotic covariance matrix obtained by the multivariate version of the δ-method (see e.g. Bishop, Fienberg and Holland (1975) [10]

(pp. 487-488)). If $\boldsymbol{f}_i = \boldsymbol{A}_i\boldsymbol{p}_i$ as above, then $\boldsymbol{P}_i = \boldsymbol{A}_i$. A consistent estimator of $\boldsymbol{H}_i$ is found by replacing $\boldsymbol{\pi}_i$ by $\boldsymbol{p}_i$.

4.2 The Linear Model and Inference

Typically one is interested in p time effects for the s treatments or subpopulations, measurements occurring at time points $t_1, t_2, \ldots, t_p$. In this case $u_i = p$ for all $i = 1, \ldots, s$. We will assume that $u_i = u$. Koch *et al.* (1977) [66] consider a linear model $\mathrm{E}(\boldsymbol{f}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$, where $\boldsymbol{X}_i$ is a known $u \times w$ design matrix, $w \leq us$, and $\boldsymbol{\beta}$ is an unknown parameter vector. If $\boldsymbol{f} = (\boldsymbol{f}_1' \; \boldsymbol{f}_2' \cdots \boldsymbol{f}_s')'$ is the full $us \times 1$ vector over subpopulations we have

$$\mathrm{E}(\boldsymbol{f}) = \boldsymbol{X}\boldsymbol{\beta} \quad \text{and} \quad \mathrm{var}(\boldsymbol{f}) = \boldsymbol{H}, \tag{4.3}$$

where $\boldsymbol{X}$ is formed by stacking the $\boldsymbol{X}_i$ and $\boldsymbol{H}$ is a block diagonal matrix with blocks $\boldsymbol{H}_i$ given by (4.2). In many situations the entries for each vector of functions $\boldsymbol{f}_i$, $(i = 1, \ldots, s)$ correspond to the p time points in each treatment or sub-population.

The parameter vector $\boldsymbol{\beta}$ is estimated by generalized least squares (GLS) with weights given by $\hat{\boldsymbol{H}}^{-1}$, the consistent estimator of $\boldsymbol{H}^{-1}$, that is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\hat{\boldsymbol{H}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{H}}^{-1}\boldsymbol{f} \tag{4.4}$$

which is a best asymptotically normal (BAN) estimator of $\boldsymbol{\beta}$. It is noted here that this method has been referred to in the literature as a WLS method, technically it a GLS method.

For complete data, $\mathbf{f}$ can be written as an $s \times p$ matrix $\mathbf{F}$, with the $\boldsymbol{f}_i'$ forming the rows. Let $\boldsymbol{M}_i$ be a known $p \times (q_i + 1)$ matrix which provides the design matrix for modelling the ith profile (using the sum of profiles model terminology of §2.6). For example, if polynomials are used, for subpopulation i, the jth row of $\boldsymbol{M}_i$ would be $(1 \; t_j \; t_j^2 \cdots t_j^{q_i})$, $j = 1, 2, \ldots, p$ and q_i is the order of the polynomial model. Consider

the model

$$\mathrm{E}(\boldsymbol{F}) = \mathrm{E}(\boldsymbol{f}_1\ \boldsymbol{f}_2 \cdots \boldsymbol{f}_s)' = \boldsymbol{D} \begin{bmatrix} \boldsymbol{\theta}_1' \boldsymbol{M}_1' \\ \boldsymbol{\theta}_2' \boldsymbol{M}_2' \\ \vdots \\ \boldsymbol{\theta}_v' \boldsymbol{M}_v' \end{bmatrix} \tag{4.5}$$

analogous to (2.24), where $\boldsymbol{D}$ is an $s \times v$ full-rank design matrix ($v \leq s$) which relates growth across the subpopulations. If $\boldsymbol{D} = [\boldsymbol{D}_1\ \boldsymbol{D}_2 \cdots \boldsymbol{D}_v]$ then we can write (4.5) as a sum of profiles model, that is

$$\mathrm{E}(\boldsymbol{F}) = \sum_{j=1}^{v} \boldsymbol{D}_j \boldsymbol{\theta}_j' \boldsymbol{M}_j'.$$

Equation (4.5) is of the form discussed by Verbyla and Cullis (1990) [149] in the continuous case. In vector form we have (4.3) with linear model (2.23), where $m = \nu$.

If the $\boldsymbol{M}_i = \boldsymbol{M}$, that is the profile models for all the subpopulations are the same, then (4.5) reduces to equation (2.2) in Stanek and Diehl (1988) [134] (apart from different notation), namely the profile model

$$\mathrm{E}(\boldsymbol{F}) = \boldsymbol{D}[\boldsymbol{\theta}_1\ \boldsymbol{\theta}_2 \cdots \boldsymbol{\theta}_v]' \boldsymbol{M}' = \boldsymbol{D}\boldsymbol{\Theta}\boldsymbol{M}'. \tag{4.6}$$

In the case of incomplete data we can follow the approach of Koch *et al.* (1972) [65]. Thus for treatment i, we suppose there are m_i different response patterns due to missing data with n_{ji} units with jth response pattern, $j = 1, 2, \ldots, m_i$, and let $\boldsymbol{Y}_{ji}$ be the vector of observed frequencies. For example consider a binary response measured at 3 time points. Table 4.1 demonstrates the possible response patterns when missing data has occurred. The values n_{jik} correspond to the number observed in the k^{th} response profile, with the j^{th} response pattern and in the i^{th} subpopulation. Table 3 in Koch *et al.* (1972) [65] displays the observed response patterns for a simulated incomplete split-plot experiment with binary responses and 3 time points.

As above $\boldsymbol{Y}_{ji}$ follows a multinomial distribution. If $\boldsymbol{p}_{ji} = \boldsymbol{Y}_{ji}/n_{ji}$ and $\boldsymbol{f}_{ji} = [f_i(\boldsymbol{p}_{ji})]$, we assume that for some design matrix $\boldsymbol{X}_{ji}$, $\mathrm{E}(\boldsymbol{f}_{ji}) = \boldsymbol{X}_{ji}\boldsymbol{\beta}$. Furthermore the covariance matrix can be defined using (4.1) and (4.2) by adding the additional subscript

Table 4.1:
Response profiles with $r = 2^3$ and missing data for subpopulation i.
('-' denotes missing values)

n_{ji}	Response Profiles			n_{jik}
	1	1	1	n_{1i1}
	1	1	0	n_{1i2}
	1	0	1	n_{1i3}
n_{1i}	0	1	1	n_{1i4}
	1	0	0	n_{1i5}
	0	1	0	n_{1i6}
	0	0	1	n_{1i7}
	0	0	0	n_{1i8}
	1	1	-	n_{2i1}
n_{2i}	1	0	-	n_{2i2}
	0	1	-	n_{2i3}
	0	0	-	n_{2i4}
	1	-	1	n_{3i1}
n_{3i}	1	-	0	n_{3i2}
	0	-	1	n_{3i3}
	0	-	0	n_{3i4}
	-	1	1	n_{4i1}
n_{4i}	-	1	0	n_{4i2}
	-	0	1	n_{4i3}
	-	0	0	n_{4i4}
n_{5i}	1	-	-	n_{5i1}
	0	-	-	n_{5i2}
n_{6i}	-	1	-	n_{6i1}
	-	0	-	n_{6i2}
n_{7i}	-	-	1	n_{7i1}
	-	-	0	n_{7i2}

j. Stacking all the $\boldsymbol{f}_{ji}$ we again obtain (4.3). The analysis can now be carried out as before.

Joint estimation of the parameters can be carried out using GLS. Having estimated $\hat{\boldsymbol{\beta}}$ by (4.4), a consistent estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$ is $\hat{\boldsymbol{\Sigma}} = (\boldsymbol{X}'\hat{\boldsymbol{H}}^{-1}\boldsymbol{X})^{-1}$ and so a GLS estimate of $\boldsymbol{\theta}$, $(\boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\theta})$, is

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{A}'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\beta} \tag{4.7}$$

with variance matrix consistently estimated by $(\boldsymbol{A}'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{A})^{-1}$. The vector of differentiable functions $\mathbf{f}$ is estimated by $\hat{\boldsymbol{f}} = \boldsymbol{X}\boldsymbol{A}\hat{\boldsymbol{\theta}}$ with a consistent estimator of the variance of $\hat{\boldsymbol{f}}$ being $\boldsymbol{X}\boldsymbol{A}(\boldsymbol{A}'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{X}'$.

The Wald statistic $Q = (\boldsymbol{f} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'\hat{\boldsymbol{H}}^{-1}(\boldsymbol{f} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ can be used to test the goodness of fit of (4.3), and Q has an approximate χ^2 distribution with d.f. $= u - w$ if the sample sizes n_i are sufficiently large. Stanek and Diehl (1988) [134] also use Wald statistics to test hypotheses of the form $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ where $\mathbf{C}$ is a known $c \times w$ matrix of full rank. In particular, they consider polynomial models and discuss the determination of the appropriate order.

4.3 Order Selection of Polynomial Models

Assume the profile design matrices $\boldsymbol{M}_i$ are all identical $(\boldsymbol{M}_i = \boldsymbol{M})$ say, that is the polynomial models for all the subpopulations have the same order. Suppose that the p repeated measurements correspond to q time points measured under d conditions, that is $p = dq$. Let the order of the polynomial model be $q - 1$, that is $\boldsymbol{M}$ is a $q \times q$ matrix of natural polynomials. Then (4.5) can be written as (4.6).

Determination of the appropriate orders for the separate polynomial models of (4.5) can be achieved by a simple modification of the method used in Stanek and Diehl (1988) [134]. Assume that $\boldsymbol{D} = \boldsymbol{I}_s$ as in Stanek and Diehl (1988) [134].

Stanek and Diehl (1988) [134] transform the polynomial models to orthogonal polynomials and test globally and marginally for orthogonal trends. These tests are carried

out sequentially beginning from the highest order polynomial. Stanek and Diehl (1988) [134] state that these tests are independent. This is not strictly correct because of the dependency among the elements of the response vectors $\boldsymbol{f}_i$ for each subpopulation.

For the i^{th} subpopulation, let $\boldsymbol{P}_i$ be a $q \times q$ orthogonal matrix ($\boldsymbol{P}_i'\boldsymbol{P}_i = \boldsymbol{P}_i\boldsymbol{P}_i' = \boldsymbol{I}_q$) such that $\boldsymbol{M}_i = \boldsymbol{P}_i\boldsymbol{R}_i$ for some non-singular matrix $\boldsymbol{R}_i$. Obviously, $\boldsymbol{P}_i = \boldsymbol{P}_j$, $\forall i,\ j$. Let $\boldsymbol{P}$ be the block diagonal matrix with diagonal matrix elements $\boldsymbol{P}_i$. Post multiply (4.6) by $\boldsymbol{P}$. Define the $s \times q$ matrices $\boldsymbol{\mathcal{F}}$ and $\boldsymbol{\Psi}$ by

$$\boldsymbol{\mathcal{F}} = (\boldsymbol{\mathcal{F}}_1\ \boldsymbol{\mathcal{F}}_2 \ldots\ \boldsymbol{\mathcal{F}}_q) = \begin{bmatrix} \boldsymbol{f}_1'\boldsymbol{P}_1 \\ \boldsymbol{f}_2'\boldsymbol{P}_2 \\ \vdots \\ \boldsymbol{f}_s'\boldsymbol{P}_s \end{bmatrix}$$

and

$$\boldsymbol{\Psi} = (\boldsymbol{\Psi}_1\ \boldsymbol{\Psi}_2 \ldots\ \boldsymbol{\Psi}_q) = \begin{bmatrix} \boldsymbol{\theta}_1'\boldsymbol{R}_1' \\ \boldsymbol{\theta}_2'\boldsymbol{R}_2' \\ \vdots \\ \boldsymbol{\theta}_s'\boldsymbol{R}_s' \end{bmatrix}.$$

The model now becomes $E(\boldsymbol{\mathcal{F}}) = \boldsymbol{I}_s\boldsymbol{\Psi}$.

The q column vectors $\boldsymbol{\mathcal{F}}_k$ ($k = 1,\ 2,\ \ldots,\ q$) represent q orthogonal trends for the s subpopulations. $\mathcal{F}_{ij}$ denotes the j^{th} trend for the i^{th} subpopulation. In many cases significant trends consist of a lower-order polynomial, for example for subpopulation i, order $q_i < q$. The i^{th} row of $\boldsymbol{\mathcal{F}}$, which corresponds to subpopulation i, can be written as $E(\boldsymbol{\mathcal{F}}_i) = (\mathcal{F}_{i1}\ \mathcal{F}_{i2} \ldots\ \mathcal{F}_{iq}) = \boldsymbol{\theta}_i'\boldsymbol{R}_i'$. A consistent estimator of the covariance of $\boldsymbol{f}_i'\boldsymbol{P}_i$ is

$$\boldsymbol{V}_{\boldsymbol{\mathcal{F}}_i} = \boldsymbol{P}_i'\boldsymbol{H}_i\boldsymbol{P}_i.$$

Important trends for subpopulation i are identified by testing hypotheses of the form $E_A(\mathcal{F}_{ij}) = 0$, where E_A is the asymptotic expectation of $\mathcal{F}_{ij}$. The Wald test statistics are $Q_{ij} = \mathcal{F}_{ij}^2/V_{\mathcal{F}_{ij}}$ where $V_{\mathcal{F}_{ij}}$ is the j^{th} entry of the main diagonal of $\boldsymbol{V}_{\boldsymbol{\mathcal{F}}_i}$. Q_{ij} has an approximate χ^2 distribution with 1 d.f. under the null hypothesis. Thus for each subpopulation we can test sequentially for the appropriate degree of the polynomial.

The above development can easily be generalised to non-polynomial models.

4.4 Oviposition of flies example

Stanek and Diehl (1988) [134] consider a study of host fruit acceptance for oviposition by Rhagoletis pomonella adult female flies. The flies originated as larvae from apple or hawthorn fruit and were exposed to apple and hawthorn fruit (separately) at ages 8-9, 11-12, 15-16 and 18-19 days after adult eclosion. If the fly attempted oviposition during exposure then this was recorded as an "accept" otherwise it was recorded as a "reject". The subpopulations are the two larval origins and $p = 8$ (2 fruit $\times$ 4 ages). Of the total of 70 flies observed, 37 belonged to the apple larval origin subpopulation, the remainder belonging to the hawthorn larval origin subpopulation. Complete data was recorded on all flies. Only 30 response profiles of the possible 2^8 were recorded (see Table 1, Stanek and Diehl, 1988 [134]). Linear functions of the response that represent the proportion of accepter responses for each test fruit and age in each subpopulation are of interest; note there is an error in Table 2 of Stanek and Diehl (1988) [134] for the apple larval origin on Hawthorn test fruit at age 11-12 days and in subsequent tests of hypotheses.

One approach is to fit the suggested model and test its fit. The other approach is the sequential procedure of Stanek and Diehl (1988) [134] or its modified form suggested above.

The modified sequential approach for polynomial order is given in Table 4.2. The linear effect for the subpopulation larval apple with test fruit apple is only just non-significant leaving only the constant effect as significant. Examination of the observed values listed in Table 4.3 for larval origin and test fruit apple implies a linear effect is not inappropriate (given that the test for the linear effect was close to significance). Thus it was decided to select a polynomial model of order 1 (straight line) for larval origin and test fruit apple.

In the case of larval origin apple and test fruit hawthorn, the linear term is significant and hence a straight line adequately represents the data. For larval origin hawthorn and

Table 4.2:
Sequential Tests for Polynomial Order

		Order			
Larvae	Test Fruit	Cubic	Quad	Linear	Const.
Apple	Apple	0.45	0.54	3.68	24.45
Apple	Hawthorn	1.61	0.07	4.99	-
Hawthorn	Apple	0.40	0.34	0.05	3.94
Hawthorn	Hawthorn	3.81	0.06	12.39	-

test fruit apple, Table 4.2 implies a constant effects model, which appears clear from the observed data. Finally for larval origin hawthorn and test fruit hawthorn, the cubic term is just non-significant, the quadratic term is non-significant while the linear term is highly significant. We consider the straight line for reasons of parsimony.

We differ with Stanek and Diehl (1988) [134] in the selection of the significant polynomial effects (that is a linear one) only for the larval hawthorn and test fruit apple. This is a consequence of Stanek and Diehl's (1988) [134] method which forces the order of polynomials to be the same for all subpopulations.

Stanek and Diehl (1988) [134] model the constant and linear trends by a simple modular model with main effects for larval origin and test fruit, and a larval origin by test fruit interaction. The larval origin by test fruit interaction for the linear trend was found to be significant, and by inspection it was concluded that the interaction was due to a differential response of hawthorn flies on apple test fruit. By a separate regression model approach, interpretation of the data is simplified.

. The parameters of our final model are estimated and hence predicted values of the proportion that accept fruit are calculated. The column denoted by Predicted Values

Table 4.3:

Observed and predicted proportion (and standard errors) that accept fruit.

Larval Origin	Test Fruit	Age in Days	Observed Value (SE)	Predicted Values (SE)		
				(a)	(b)	(c)
A	A	8-9	.162 (.061)	.118 (.049)	.167 (.058)	.138 (.055)
		11-12	.216 (.068)	.186 (.043)	.209 (.050)	.202 (.049)
		15-16	.324 (.077)	.278 (.044)	.266 (.054)	.287 (.048)
		18-19	.297 (.075)	.347 (.051)	.308 (.067)	.351 (.054)
A	H	8-9	.703 (.075)	.667 (.048)	.693 (.060)	.681 (.047)
		11-12	.676 (.077)	.736 (.039)	.751 (.043)	.745 (.039)
		15-16	.865 (.056)	.827 (.036)	.829 (.038)	.831 (.037)
		18-19	.865 (.056)	.896 (.042)	.888 (.051)	.895 (.043)
H	A	8-9	.061 (.042)	.088 (.038)	.073 (.034)	.078 (.034)
		11-12	.091 (.050)	.083 (.033)	.073 (.034)	.078 (.034)
		15-16	.061 (.042)	.076 (.036)	.073 (.034)	.078 (.034)
		18-19	.061 (.042)	.071 (.043)	.073 (.034)	.078 (.034)
H	H	8-9	.455 (.087)	.514 (.058)	.506 (.075)	.572 (.058)
		11-12	.636 (.084)	.583 (.049)	.603 (.057)	.637 (.052)
		15-16	.636 (.084)	.674 (.046)	.732 (.050)	.722 (.050)
		18-19	.849 (.062)	.743 (.049)	.829 (.061)	.786 (.053)

(a) in Table 4.3 represents the predicted values of Table 2 in Stanek and Diehl (1988) [134]. The predicted values for the model chosen here (linear models except for larval origin hawthorn and test fruit apple which is constant) are listed in Table 4.3 under (b). Parameters are estimated by the GLS equation (4.7), where $\hat{\boldsymbol{\beta}} = \boldsymbol{f}$ from (4.4) as $\boldsymbol{X}$ is the identity matrix. A correlation of 1.00 for hawthorn larval origin flies tested on apple fruit at 15-16 and 18-19 days presents a problem in that this implies the weight matrix $\hat{\boldsymbol{\Sigma}}$ in (4.7) is singular. To overcome this problem $\hat{\boldsymbol{\Sigma}}$ is replaced by a generalized inverse matrix (see Searle (1971) [128] (pp. 1-30)).

Stanek and Diehl (1988) [134] test and retain the hypothesis of parallel slopes among flies on all test fruit except hawthorn flies on apple fruit. As there is no linear term for hawthorn flies on apple fruit, we test parallel slopes for the other fly, test fruit combinations. We retain the hypothesis of parallel slopes, with the Wald statistic $Q = 2.2003$ on 2 degrees of freedom. The predicted values found by fitting this model, are recorded in Table 4.3 under (c).

Predicted values under (a), (b) and (c) are reasonable compared to the observed values given the small sample size. When compared to (a) and (c), nine of the sixteen predicted values under (b) are the closest to the observed values. This is not surprising because of the restriction of parallel slope under (a) and (c). Differences between (a) and (c) are generally not very large. The observed values however for hawthorn-apple indicates a constant model may be more suitable. This is supported by the sequential tests of Table 4.2.

4.5 Discussion

The approach presented in this chapter extends the work of Stanek and Diehl (1988) [134] in a straightforward manner. Connections are made with the continuous case especially as presented by Verbyla and Cullis (1990) [149] in terms of sum of profiles. In modelling the subpopulations separately we have incorporated a greater level of flexibility rather than rigidly forcing all subpopulations to have polynomials of the same order.

Although polynomials are used, following the work of Verbyla and Cullis (1990) [149], and Verbyla and Venables (1988) [151], other model forms could be used.

Chapter 5

Quasi-likelihood and Generalized Estimating Equations

5.1 Quasi-likelihood

Let $\boldsymbol{y}$ be an $n \times 1$ response vector of n independent units with mean vector $\boldsymbol{\mu}$ and covariance matrix $\phi \boldsymbol{V}(\boldsymbol{\mu})$. The matrix $\boldsymbol{V}(\boldsymbol{\mu})$ is a matrix of known functions,

$$\boldsymbol{V}(\boldsymbol{\mu}) = \text{diag}(g_1(\mu_1),\ g_2(\mu_2), \ldots,\ g_n(\mu_n))$$

where the function $g_j(\mu_j)$ specifies the mean and variance relationships. It is usually assumed that $g_j(\mu_j) = g(\mu_j)$, though this is not necessary for the material in this section. The mean is related to the linear predictor, $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$, by $\mu_i = h(\boldsymbol{x}_i'\boldsymbol{\beta})$ where $\boldsymbol{\beta}$ is an $m \times 1$ vector of parameters, and $\boldsymbol{x}_i$ is a vector of explanatory variables or covariates. The inverse of h is the link function.

The function, for a single observation, y_i say,

$$U_i = U(\mu_i; y_i) = \frac{y_i - \mu_i}{\phi g(\mu_i)}$$

has the following properties,

(a) $E(U_i) = 0$

(b) $\text{var}(U_i) = 1/\phi g(\mu_i)$

(c) $-E(\partial U_i/\partial \mu_i) = 1/\phi g(\mu_i)$.

These 3 properties are shared by the log-likelihood derivative, and are important for most first-order asymptotic theory for likelihood functions. The integral

$$Q(\mu_i; y_i) = \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi g(\mu_i)} dt,$$

if it exists, behaves in a similar manner to a likelihood function for μ_i. It is referred to as a quasi-likelihood for μ_i based on y_i. It is more correct to refer to $Q(\mu_i; y_i)$ as a log quasi-likelihood function. Due to the independence of the elements of $\boldsymbol{y}$, the quasi-likelihood for complete data is

$$Q(\boldsymbol{\mu}; \boldsymbol{y}) = \sum_{i=1}^{n} Q(\mu_i; y_i).$$

Similarly the function

$$\mathcal{D}(\boldsymbol{y}; \boldsymbol{\mu}) = -2\phi \sum_{i=1}^{n} Q(\mu_i; y_i)$$

is the analog of the deviance function, and is referred to as the quasi-deviance function.

By differentiating $Q(\boldsymbol{\mu}; \boldsymbol{y})$ with respect to β_k, the quasi-likelihood estimating equations for the regression parameters $\boldsymbol{\beta}$ is

$$\frac{\partial Q}{\partial \beta_k} = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta_k} g(\mu_i)^{-1}(y_i - \mu_i)/\phi = 0, \quad k = 1, \ldots, m,$$

which in vector form is

$$U(\boldsymbol{\beta}) = \boldsymbol{D}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})/\phi \tag{5.1}$$

where $\boldsymbol{D}$ is an $n \times m$ matrix with components $D_{ik} = \partial \mu_i/\partial \beta_k$, and $\boldsymbol{V}$ is the $m \times m$ diagonal matrix with diagonal elements $g(\mu_i)$. The estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, is obtained by solving (5.1). Equation (5.1) is referred to as the quasi-score function. The covariance matrix of $U(\boldsymbol{\beta})$ is

$$\text{var}(U(\boldsymbol{\beta})) = -E\left(\frac{\partial U(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right) = \boldsymbol{i}_{\boldsymbol{\beta}} = \boldsymbol{D}'\boldsymbol{V}^{-1}\boldsymbol{D}/\phi.$$

The asymptotic covariance of $\hat{\boldsymbol{\beta}}$ is

$$\text{var}(\hat{\boldsymbol{\beta}}) = \boldsymbol{i}_{\boldsymbol{\beta}}^{-1} = \phi(\boldsymbol{D}'\boldsymbol{V}^{-1}\boldsymbol{D})^{-1}$$

analogous to Fisher's information matrix.

Estimation of $\hat{\boldsymbol{\beta}}$ can be performed by the Newton-Raphson method with Fisher Scoring, that is

$$\hat{\boldsymbol{\beta}}_j = \hat{\boldsymbol{\beta}}_{j-1} + (\hat{\boldsymbol{D}}'_{j-1}\hat{\boldsymbol{V}}_{j-1}^{-1}\hat{\boldsymbol{D}}_{j-1})^{-1}\hat{\boldsymbol{D}}_{j-1}\hat{\boldsymbol{V}}_{j-1}^{-1}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_{j-1})$$

for the j^{th} iteration, and $\hat{\boldsymbol{D}}_{j-1}$, $\hat{\boldsymbol{V}}_{j-1}^{-1}$ and $\hat{\boldsymbol{\mu}}_{j-1}$ are $\boldsymbol{D}$, $\boldsymbol{V}$ and $\boldsymbol{\mu}$ evaluated at $\hat{\boldsymbol{\beta}}_{j-1}$. The conventional estimate of ϕ is the moment estimator

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2/g_i(\hat{\mu}_i).$$

The estimate $\hat{\boldsymbol{\beta}}$ is asymptotically unbiased and Normally distributed.

5.2 Extended Quasi-likelihood

5.2.1 Specifying the Extended Quasi-likelihood

To include analysis and estimation of the dispersion parameter and the variance function, extended quasi-likelihood was introduced by Nelder and Pregibon (1987) [100]. This section is included because modelling the dispersion parameter with GEE for longitudinal data is examined in chapter 6. The extended quasi-likelihood approach and the related work of Smyth (1989) [131] provides the impetus for chapter 6.

The extended quasi-likelihood for a single observation y_i, with mean μ_i and variance $\phi g(\mu_i)$ is

$$Q^+(\mu_i, \phi; y_i) = -\frac{\mathcal{D}(y_i; \mu_i)}{2\phi} - \frac{\log(\phi)}{2}, \tag{5.2}$$

allowing Q^+ to act as a log-likelihood with respect to ϕ because $E(\partial Q^+/\partial\phi) = 0$. See Nelder and Pregibon (1987) [100] and McCullagh and Nelder (1989) [92] (pp. 349-351) for a more detailed discussion. The second term on the right hand side of (5.2) is replaced

by $-\log(2\pi\phi g(\mu_i))/2$ in Nelder and Pregibon (1987) [100]. Since we are assuming that the y_i are independent then

$$Q^+(\boldsymbol{\mu}, \phi; \boldsymbol{y}) = \sum_{i=1}^{n} Q^+(\mu_i, \phi; y_i).$$

If we assume that ϕ is sufficiently small so that the approximation $E(\mathcal{D}(y_i; \mu_i)) = \phi$ is reasonable, then the derivatives

$$\frac{\partial Q^+}{\partial \mu_i} = \frac{y_i - \mu_i}{\phi g(\mu_i)} \quad \text{and} \quad \frac{\partial Q^+}{\partial \phi} = \frac{\mathcal{D}(y_i; \mu_i)}{2\phi^2} - \frac{1}{2\phi}$$

have zero mean, and consequently the estimate of $\boldsymbol{\beta}$ obtained by maximizing Q^+ is the same as maximizing Q. The estimate of the dispersion parameter ϕ obtained by maximizing Q^+ is $\hat{\phi} = \mathcal{D}(\boldsymbol{y}; \hat{\boldsymbol{\mu}})/n$. When the underlying distribution is the Normal or inverse Gaussian then $\hat{\phi}$ is the maximum likelihood estimate of ϕ.

5.2.2 Indexed Variance Functions

We can relax the requirement that the variance of y_i is known up to a multiplicative constant, ϕ. Suppose the variance function is an unknown parameter ζ. Thus we can write the variance of y_i as $\text{var}(y_i) = \phi g_\zeta(\mu_i)$. This implies for a single observation the extended quasi-likelihood is

$$Q_\zeta^+(y_i; \mu_i) = -\frac{\mathcal{D}_\zeta(y_i; \mu_i)}{2\phi} - \frac{\log(\phi)}{2}$$

where

$$\mathcal{D}_\zeta(y_i; \mu_i) = -2\int_{y_i}^{\mu_i} \frac{y_i - u_i}{g_\zeta(u_i)} du_i.$$

An important example for $g_\zeta(\mu_i)$ is $g_\zeta(\mu_i) = \mu_i^\zeta$, with commonly occurring values of ζ being 0, 1, 2 and 3. These values correspond to variance functions from the Normal, Poisson, Gamma and inverse Gaussian distributions respectively.

5.2.3 A Variable Dispersion Parameter

Suppose the dispersion parameter varies across the observations such that

$$\text{var}(y_i) = \phi_i g(\mu_i),$$

and is assumed to be a function of a set of covariates,

$$b(\phi_i) = \boldsymbol{a}_i'\boldsymbol{\gamma}$$

where b is a link function, $\boldsymbol{a}_i$ is a $k \times 1$ vector of covariates and $\boldsymbol{\gamma}$ is the corresponding vector of unknown parameters. The covariate vector $\boldsymbol{a}_i$ may be a subset of $\boldsymbol{x}_i$. A desired goal of the dispersion model is to improve estimation of the mean parameters (lower standard errors).

Let $d_i(y_i;\mu_i)$ be a suitably chosen statistic such that

$$E(d_i(y_i;\mu_i)) = \phi_i \quad \text{and} \quad \text{var}(d_i(y_i;\mu_i)) = \tau v_d(\phi_i),$$

where $v_d(\phi_i)$ is the dispersion variance function. McCullagh and Nelder (1989) [92] (pp. 358-359) consider two possible choices for the dispersion statistic d_i,

(a) Generalized Pearson statistic,

$$d_i(y_i;\mu_i) = r_{pi}^2 = (y_i - \mu_i)^2/g(\mu_i),$$

(b) Deviance statistic for the i^{th} unit,

$$d_i(y_i;\mu_i) = r_{di}^2 = 2\int_{\mu_i}^{y_i} \frac{y_i - t}{g(t)}dt.$$

Using Normal theory, the two statistics r_{pi}^2 and r_{di}^2 are equivalent; this is not true when Normality does not hold. Though $E(r_{pi}^2) = \phi_i$ exactly, the expected value of the deviance statistic r_{di}^2 is only approximately ϕ_i. If y_i is Normal, d_i has the $\phi_i\chi_1^2$ distribution and consequently we would choose the Gamma model with $v(\phi_i) = 2\phi_i^2$. A second quasi-likelihood equation for the d_i's will lead to an estimate for $\boldsymbol{\gamma}$.

The second quasi-likelihood can be written as

$$-2Q^+ = \sum_{i=1}^{n} \frac{d_i}{\phi_i} + \sum_{i=1}^{n} \log\left(2\pi\phi_i g(y_i)\right)$$

where $d_i = r_{di}$. The estimating equations for $\boldsymbol{\beta}$ are

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi_i g(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0, \; j = 1, \ldots, \; m, \tag{5.3}$$

which are the quasi-likelihood equations except $1/\phi_i$ is included as a weight.

The estimating equations for γ are

$$\sum_{i=1}^{n} \frac{d_i - \phi_i}{\phi_i^2} \frac{\partial \phi_i}{\partial \gamma_j} = 0, \; j = 1, \ldots, \; k. \tag{5.4}$$

Equations (5.3) and (5.4) are the quasi-likelihood with deviance components $d_i = r_{di}^2$. Often, however the variance of d_i exceeds $2\phi_i^2$ and an appropriate adjustment should be considered (see McCullagh and Nelder (1989) [92] (pp. 361-362)).

The difficulty with using $d_i = r_{di}^2$ or equivalently the extended quasi-likelihood, is because, as discussed earlier, $E(r_{di}^2)$ is not exactly ϕ_i. It is not generally possible to derive the expectations of r_{di}^2 without full distributional results.

Smyth (1989) [131] obtains estimates of the mean and dispersion parameters by quasi-likelihood equations of the form (5.3) and (5.4), but uses the generalized Pearson residuals $d_i = r_{pi}^2$ instead. Further, Smyth (1989) [131] observes that $\boldsymbol{\beta}$ and γ are quasi-orthogonal by differentiation of (5.3) by γ_j and (5.4) by β_k and taking expectations.

5.3 Marginal Means and First Order Generalized Estimating Equations

5.3.1 Extension of the Quasi-likelihood Equations for Longitudinal Data

Liang and Zeger (1986) [81] proposed an extension of generalized linear models for longitudinal data. A class of estimating equations was introduced that produce, under mild regularity conditions, consistent estimates of the regression parameters. The estimating equations they consider attempt to account for the correlations among the repeated measurements for a given unit. The distribution of the response values is presumed to belong to the exponential family.

The joint distribution of a unit's observation is not specified, but the estimating

equations reduce to the score equations for Normally distributed multivariate data. The estimating equations can be thought of as an extension of the quasi-likelihood equations.

Zeger and Liang (1986) [167] propose, in a more general way, methodology for discrete and continuous data that is based on quasi-likelihood theory. This extension of the quasi-likelihood case to longitudinal data makes no specification of the underlying marginal distribution for the marginal responses for each unit. Only first and second moment assumptions are made. The resulting estimating equations are referred to as generalized estimating equations (GEE).

5.3.2 GEE for the Mean Parameters

Suppose we have observations $(y_{ij},\ \boldsymbol{x}_{ij})$ for subjects $i = 1,\ldots,n$ at time points t_{ij}, $j = 1,\ldots,p_i$, where $\boldsymbol{x}_{ij}$ is an $m \times 1$ vector of explanatory variables that may be time-dependent or time-independent or a mixture of both. Let μ_{ij} be the expectation of y_{ij}, $\boldsymbol{y}_i$ and $\boldsymbol{\mu}_i$ be the $p_i \times 1$ vectors $(y_{i1},\ldots,y_{ip_i})'$ and $(\mu_{i1},\ldots,\mu_{ip_i})'$ respectively. One of the assumptions made is that the mean is related to the linear predictor $\eta_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta}$ by

$$\mu_{ij} = h(\boldsymbol{x}_{ij}'\boldsymbol{\beta}), \tag{5.5}$$

where $\boldsymbol{\beta}$ is an $m \times 1$ vector of parameters and the inverse of h is the link function. Further, it is assumed that the variance of y_{ij} is

$$\text{var}(y_{ij}) = v_{ij} = \phi g(\mu_{ij}) \tag{5.6}$$

where ϕ is a scale parameter. Interest lies in $\boldsymbol{\beta}$, and ϕ is treated as a nuisance parameter.

To account for the correlations within a subject let $\boldsymbol{R}_i(\boldsymbol{\alpha})$ represent a $p_i \times p_i$ 'working' correlation matrix of $\boldsymbol{y}_i$, so named because we do not expect it to be correctly specified. $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is assumed to be fully specified by an $s \times 1$ vector of unknown parameters, $\boldsymbol{\alpha}$, which is the same for all subjects even though $\boldsymbol{R}_i(\boldsymbol{\alpha})$ may differ from subject to subject. A working covariance matrix for $\boldsymbol{y}_i$ is given by

$$\boldsymbol{V}_{bi} = \phi \boldsymbol{A}_i^{1/2} \boldsymbol{R}_i(\boldsymbol{\alpha}) \boldsymbol{A}_i^{1/2},$$

where $\boldsymbol{A}_i$ is a $p_i \times p_i$ diagonal matrix with $g(\mu_{ij})$ as the jth diagonal element. The GEE's for longitudinal data are

$$\sum_{i=1}^{n} \boldsymbol{U}_{bi}(\boldsymbol{\beta},\ \boldsymbol{\alpha}) = \sum_{i=1}^{n} \boldsymbol{D}_{bi}' \boldsymbol{V}_{bi}^{-1} \boldsymbol{S}_{bi} = \mathbf{0}, \tag{5.7}$$

where $\boldsymbol{S}_{bi} = \boldsymbol{y}_i - \boldsymbol{\mu}_i$ and $\boldsymbol{D}_{bi} = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}$. When $p_i = 1$ for all i, (5.7) reduces to the quasi-likelihood equations (5.1). When Normality is assumed, $\sum_{i=1}^{n} \boldsymbol{D}_{bi}' \boldsymbol{V}_{bi}^{-1} \boldsymbol{S}_{bi}$ is the score vector of $\boldsymbol{\beta}$.

The estimating equations (5.7) depend on $\boldsymbol{\alpha}$ as well as $\boldsymbol{\beta}$. If $\boldsymbol{\alpha}$ is replaced by an $n^{1/2}$-consistent estimator $\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta},\ \phi)$ and ϕ by an $n^{1/2}$-consistent estimator $\hat{\phi}(\boldsymbol{\beta})$, then (5.7) becomes a function of $\boldsymbol{\beta}$ alone, that is

$$\sum_{i=1}^{n} \boldsymbol{U}_{bi}(\boldsymbol{\beta},\ \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta},\ \hat{\phi}(\boldsymbol{\beta}))) = \mathbf{0}. \tag{5.8}$$

The following definition is used in Theorem 5.1.

Definition 5.1 *Consider a sequence of random variables $\{a_n\}$. Then $O_p(a_n)$ denotes a random variable such that for all $\epsilon > 0$, there is a constant K and an integer n_0 such that*

$$P(|O_p(a_n)/a_n| < K) > 1 - \epsilon \quad \forall\ n > n_0,$$

where P signifies the probability.

For a given $\boldsymbol{R}_i(\boldsymbol{\alpha})$, let the solution of (5.8) be $\hat{\boldsymbol{\beta}}_R$. The following theorem applies to $\hat{\boldsymbol{\beta}}_R$.

Theorem 5.1 *Under mild regularity conditions and given that:*

(i) $\hat{\boldsymbol{\alpha}}$ is $n^{1/2}$-consistent given $\boldsymbol{\beta}$ and ϕ,

(ii) $\hat{\phi}$ is $n^{1/2}$-consistent given $\boldsymbol{\beta}$ and

(iii) $|\partial \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta},\ \phi)/\partial \phi| \leq H(\boldsymbol{Y},\ \boldsymbol{\beta})$ which is $O_p(1)$,

then $n^{1/2}(\hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta})$ is asymptotically multivariate Normal with zero expectation and covariance matrix

$$\boldsymbol{V}_R = \lim_{n \to \infty} n(\boldsymbol{G}_1^{-1} \boldsymbol{G}_0 \boldsymbol{G}_1^{-1}), \tag{5.9}$$

where

$$G_0 = \sum_{i=1}^{n} D_{bi}' V_{bi}^{-1} \mathrm{cov}(\boldsymbol{y}_i) V_{bi}^{-1} D_{bi} \quad \textit{and} \quad G_1 = \sum_{i=1}^{n} D_{bi}' V_{bi}^{-1} D_{bi}.$$

Proof: A sketch of the proof is given in the Appendix of Liang and Zeger (1986) [81].

The asymptotic covariance matrix $\boldsymbol{V}_R$ can be estimated by replacing $\mathrm{cov}(\boldsymbol{y}_i)$ by $\boldsymbol{S}_{bi}\boldsymbol{S}_{bi}'$, and $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and ϕ by their estimates. This estimate will be denoted as $\hat{\boldsymbol{V}}_R$. The vector $\hat{\boldsymbol{\beta}}_R$ and matrix $\hat{\boldsymbol{V}}_R$ are consistent estimators of $\boldsymbol{\beta}$ and $\boldsymbol{V}_R$ respectively if the mean is correctly specified. Consistency does not depend on the correct selection of $R_i(\alpha)$ however.

To compute $\hat{\boldsymbol{\beta}}_R$, Liang and Zeger (1986) [81] use an iteration scheme based on a modified Fisher scoring method for $\boldsymbol{\beta}$ and moment estimation of the nuisance parameters $\boldsymbol{\alpha}$ and ϕ. At iteration c we have, given the current estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{\phi}$,

$$\hat{\boldsymbol{\beta}}_R^c = \hat{\boldsymbol{\beta}}_R^{c-1} - \{\boldsymbol{G}_1(\hat{\boldsymbol{\beta}}_R^{c-1}, \hat{\boldsymbol{\alpha}}^*(\hat{\boldsymbol{\beta}}_R^{c-1}))\}^{-1} \boldsymbol{U}_b(\hat{\boldsymbol{\beta}}_R^{c-1}, \hat{\boldsymbol{\alpha}}^*(\hat{\boldsymbol{\beta}}_R^{c-1})), \tag{5.10}$$

where $\hat{\boldsymbol{\alpha}}^*(\boldsymbol{\beta}) = \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta}))$ and $\boldsymbol{G}_1$ is as defined above and is the limiting value of the expectation of the derivative of $\sum_{i=1}^{n} \boldsymbol{U}_{bi}(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta})))$ and $\boldsymbol{U}_b$ is the GEE quasi-score $\sum_{i=1}^{n} \boldsymbol{D}_{bi}' \boldsymbol{V}_{bi}^{-1} \boldsymbol{S}_{bi}$. Let $\boldsymbol{D} = (\boldsymbol{D}_{b1}', \ldots, \boldsymbol{D}_{bm}')'$, $\boldsymbol{S} = (\boldsymbol{S}_{b1}', \ldots, \boldsymbol{S}_{bm}')'$ and $\boldsymbol{V}$ be a $k \times k$ block diagonal matrix, $k = \sum_{i=1}^{n} p_i$, with $\boldsymbol{V}_{bi}(\hat{\boldsymbol{\beta}}_R^{c-1}, \hat{\boldsymbol{\alpha}}^*(\hat{\boldsymbol{\beta}}_R^{c-1}))$ as the diagonal elements. If we let $\boldsymbol{Z} = \boldsymbol{D}\boldsymbol{\beta} - \boldsymbol{S}$, then (5.10) is equivalent to an iteratively re-weighted linear regression of $\boldsymbol{Z}$ on $\boldsymbol{D}$ with weight matrix $\boldsymbol{V}^{-1}$.

At a given iteration the nuisance parameters $\boldsymbol{\alpha}$ and ϕ can be estimated from the current Pearson residuals

$$\hat{r}_{ij} = (y_{ij} - \mu_{ij}(\hat{\boldsymbol{\beta}}_R))/g(\mu_{ij}(\hat{\boldsymbol{\beta}}_R))^{1/2},$$

where $\hat{\boldsymbol{\beta}}_R$ (suppressing the superscript c) is the current value of $\boldsymbol{\beta}$ in the iterative procedure. The scale parameter ϕ can be estimated by the moment estimator

$$\hat{\phi} = \sum_{i=1}^{n} \sum_{j=1}^{p_i} \hat{r}_{ij}^2/(k-m),$$

which is $n^{1/2}$-consistent. The correlation parameter $\boldsymbol{\alpha}$ is estimated consistently by borrowing strength over the n subjects. The specific estimator depends on the choice of

$\boldsymbol{R}(\boldsymbol{\alpha})$ and is usually estimated by a simple function of

$$\hat{R}_{st} = \sum_{i=1}^{n} \hat{r}_{is}\hat{r}_{it}/(n-m).$$

Examples of several different choices for $\boldsymbol{R}(\boldsymbol{\alpha})$ and their parameter estimators are given in Liang and Zeger (1986) [81]. Moment estimators for AR-1, Equal and Unstructured (fully parameterized) correlation matrices are discussed in chapter 6.

5.4 A Second Order Extension of the GEE

Prentice (1988) [109] extends the GEE's of Liang and Zeger (1986) [81] to involve a second set of estimating equations of the same form as (5.8) when examining regression methods for correlated binary data. The response vector for this second set is the vector of sample correlations $\boldsymbol{r}_i = \{r_{ist}\}$,

$$r_{ist} = r_{ist}(\boldsymbol{\beta}) = \frac{(y_{is}-\pi_{is})(y_{it}-\pi_{it})}{(\pi_{is}(1-\pi_{is})\pi_{it}(1-\pi_{it}))^{1/2}},$$

where y_{ij} is the jth binary response in block i and $\pi_{ij} = E(y_{ij}) = \mathrm{pr}(y_{ij}=1)$. The sample correlation r_{ist} has expectation $E(r_{ist}) = \mathrm{corr}(y_{is},\ y_{it}) = \rho_{ist}(\boldsymbol{\alpha})$ and variance

$$w_{ist}(\boldsymbol{\alpha}) = 1 + \frac{(1-2\pi_{is})(1-2\pi_{it})\rho_{ist}}{(\pi_{is}(1-\pi_{is})\pi_{it}(1-\pi_{it}))^{1/2}} - \rho_{ist}^2.$$

The mean response vector $\boldsymbol{\rho}_i = \{\rho_{ist}(\alpha)\}$ is modelled as a linear function of $\boldsymbol{\alpha}$. The GEE for estimation of $\boldsymbol{\alpha}$ from the response vector $\boldsymbol{r}_i$ is

$$\begin{aligned} \sum_{i=1}^{n} \boldsymbol{U}_{ri}(\boldsymbol{\beta},\ \boldsymbol{\gamma},\ \boldsymbol{\alpha}) &= \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\rho}_i}{\partial \boldsymbol{\alpha}}\right)' \boldsymbol{V}_{ri}^{-1}(\boldsymbol{r}_i - \boldsymbol{\rho}_i) \\ &= \sum_{i=1}^{n} \boldsymbol{D}_{ri}'\boldsymbol{V}_{ri}^{-1}(\boldsymbol{r}_i - \boldsymbol{\rho}_i) = \boldsymbol{0}. \end{aligned} \tag{5.11}$$

Under mild regularity conditions, Prentice (1988) [109] demonstrates that the joint asymptotic distribution of $n^{1/2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$ and $n^{1/2}(\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha})$ is Normal with mean zero and variance matrix being n times

$$\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{0} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{21}' \\ \boldsymbol{0} & \boldsymbol{A}_{22} \end{bmatrix} \tag{5.12}$$

where

$$
\begin{aligned}
\boldsymbol{A}_{11} &= (\sum_{i=1}^{n} \boldsymbol{D}'_{bi}\boldsymbol{V}_{bi}^{-1}\boldsymbol{D}_{bi})^{-1}, \\
\boldsymbol{A}_{22} &= (\sum_{i=1}^{n} \boldsymbol{D}'_{ri}\boldsymbol{V}_{ri}^{-1}\boldsymbol{D}_{ri})^{-1}, \\
\boldsymbol{A}_{21} &= \boldsymbol{A}_{22}\left(\sum_{i=1}^{n} \boldsymbol{D}'_{ri}\boldsymbol{V}_{ri}^{-1}\frac{\partial \boldsymbol{r}_i}{\partial \boldsymbol{\beta}}\right)^{-1}\boldsymbol{A}_{11}, \\
\boldsymbol{\Lambda}_{11} &= \boldsymbol{D}'_{bi}\boldsymbol{V}_{bi}^{-1}\mathrm{var}(\boldsymbol{y}_i)\boldsymbol{V}_{bi}^{-1}\boldsymbol{D}_{bi}, \\
\boldsymbol{\Lambda}_{22} &= \boldsymbol{D}'_{ri}\boldsymbol{V}_{ri}^{-1}\mathrm{var}(\boldsymbol{r}_i)\boldsymbol{V}_{ri}^{-1}\boldsymbol{D}_{ri}, \\
\boldsymbol{\Lambda}_{12} &= \boldsymbol{D}'_{bi}\boldsymbol{V}_{bi}^{-1}\mathrm{cov}(\boldsymbol{y}_i,\ \boldsymbol{r}_i)\boldsymbol{V}_{ri}^{-1}\boldsymbol{D}_{ri}, \\
\boldsymbol{\Lambda}_{21} &= \boldsymbol{\Lambda}'_{12}.
\end{aligned}
$$

Equation (5.12) may be estimated by replacing $\mathrm{var}(\boldsymbol{y}_i)$, $\mathrm{cov}(\boldsymbol{y}_i,\ \boldsymbol{r}_i)$ and $\mathrm{var}(\boldsymbol{r}_i)$ by $\boldsymbol{S}_{bi}\boldsymbol{S}'_{bi}$, $\boldsymbol{S}_{bi}(\boldsymbol{r}_i - \boldsymbol{\rho}_i)'$ and $(\boldsymbol{r}_i - \boldsymbol{\rho}_i)(\boldsymbol{r}_i - \boldsymbol{\rho}_i)'$ respectively, and evaluating $(\boldsymbol{\beta},\ \boldsymbol{\alpha})$ at $(\tilde{\boldsymbol{\beta}},\ \tilde{\boldsymbol{\alpha}})$, where $(\tilde{\boldsymbol{\beta}},\ \tilde{\boldsymbol{\alpha}})$ are the GEE estimators derived from the two sets of estimating equations.

Lipsitz *et al.* (1991) [86] recommend the use of the odds ratio as an alternative to the correlation coefficient as a measure of association between pairs of correlated binary response. They take as their response vector the $p_i(p_i - 1)$ random vector $\boldsymbol{r}_i = \{y_{ist}\}$, where $y_{ist} = y_{is}y_{it}$. The expected value of y_{ist} is the joint probability $\pi_{ist} = \mathrm{pr}(y_{is} = 1,\ y_{it} = 1)$ and so $E(\boldsymbol{r}_i) = \boldsymbol{\delta}_i = \{\pi_{ist}\}$. The joint probability π_{ist} can be written in terms of the marginal probabilities π_{is}, π_{it} and the odds ratio

$$\tau_{ist} = \tau_{ist}(\boldsymbol{\alpha},\ \boldsymbol{\beta}) = \frac{\pi_{ist}(1 - \pi_{is} - \pi_{it} - \pi_{ist})}{(\pi_{is} - \pi_{ist})(\pi_{it} - \pi_{ist})},$$

as shown in equation (6) in Lipsitz *et al.* (1991) [86]. The joint probability π_{ist} is a function of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ and the expectation of the log of the odds ratio can be modelled as a linear function of $\boldsymbol{\alpha}$.

Liang *et al.* (1992) [82] discuss the usage of the odds ratio for not only the binary case but also for the polytomous case, as well as mixed continuous/discrete responses.

Using (5.7) and (5.11) to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ separately assumes $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are orthogonal parameters (Cox and Reid (1987) [26]). Liang (1992) *et al.* [82] refer to (5.7) and (5.11) as GEE1.

Suppose that $\boldsymbol{f}_i = (\boldsymbol{y}_i', \boldsymbol{r}_i')'$, where $\boldsymbol{f}_i$ has mean vector $\boldsymbol{\eta}_i = (\boldsymbol{\mu}_i', \boldsymbol{\rho}_i')'$, and let $\boldsymbol{\lambda} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$. As we have seen, estimates of $\boldsymbol{\lambda}$ can be obtained from GEE1. We can also estimate $\boldsymbol{\lambda}$ by the set of unbiased estimating equations,

$$\boldsymbol{U}(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \boldsymbol{D}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{S}_i = \boldsymbol{0} \tag{5.13}$$

where $\boldsymbol{S}_i = \boldsymbol{f}_i - \boldsymbol{\eta}_i$, $\boldsymbol{\Sigma}_i$ is the covariance matrix of $\boldsymbol{f}_i$ and

$$\boldsymbol{D}_i = \partial \boldsymbol{\eta}_i / \partial \boldsymbol{\lambda} = \begin{bmatrix} \boldsymbol{D}_{bi} & 0 \\ \partial \boldsymbol{\rho}_i / \partial \boldsymbol{\beta} & \boldsymbol{D}_{ri} \end{bmatrix}.$$

The estimating equations (5.13) are referred to as GEE2 by Liang *et al.* (1992) [82]. Note that we can write (5.7) and (5.11) as the combined GEE

$$\sum_{i=1}^{n} \begin{bmatrix} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} & \boldsymbol{0} \\ \boldsymbol{0} & \frac{\partial \boldsymbol{\rho}_i}{\partial \boldsymbol{\alpha}} \end{bmatrix}' \begin{bmatrix} \mathrm{var}(\boldsymbol{y}_i) & \boldsymbol{0} \\ \boldsymbol{0} & \mathrm{var}(\boldsymbol{r}_i) \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{y}_i - \boldsymbol{\mu}_i \\ \boldsymbol{r}_i - \boldsymbol{\rho}_i \end{bmatrix} = \boldsymbol{0} \tag{5.14}$$

which is a more concise way of specifying GEE1.

If

$$\boldsymbol{V}_i = \begin{bmatrix} \boldsymbol{V}_{bi} & \boldsymbol{V}_{bri} \\ \cdot & \boldsymbol{V}_{ri} \end{bmatrix}$$

replaces $\boldsymbol{\Sigma}_i$ in (5.13) as a working covariance matrix, then GEE1 is equivalent to setting $\partial \boldsymbol{\rho}_i / \partial \boldsymbol{\beta} = \boldsymbol{0}$ and $\boldsymbol{V}_{bri} = \boldsymbol{0}$.

Let the $v \times 1$ parameter vector $\boldsymbol{\xi}$ completely specify the matrices $\boldsymbol{V}_{bri}$ and $\boldsymbol{V}_{ri}$ and let $\hat{\boldsymbol{\xi}}(\boldsymbol{\lambda})$ be an $n^{1/2}$-consistent estimate of $\boldsymbol{\xi}$. The vector $n^{1/2}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda})$, where $\hat{\boldsymbol{\lambda}}$ is the solution of (5.13), has as $n \to \infty$, an asymptotic multivariate normal distribution with mean zero and covariance matrix that can be consistently estimated by

$$\boldsymbol{V}_R^* = n(\boldsymbol{H}_1^{-1} \boldsymbol{H}_0 \boldsymbol{H}_1^{-1}), \tag{5.15}$$

if the mean vector is correctly specified and certain mild regularity conditions hold. The matrices $\boldsymbol{H}_0$ and $\boldsymbol{H}_1$ in (5.15) are

$$\boldsymbol{H}_0 = \sum_{i=1}^{n} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \boldsymbol{S}_i \boldsymbol{S}_i' \boldsymbol{V}_i^{-1} \boldsymbol{D}_i \quad \text{and} \quad \boldsymbol{H}_1 = \sum_{i=1}^{n} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} (\boldsymbol{D}_i + \partial \boldsymbol{f}_i / \partial \hat{\boldsymbol{\lambda}}),$$

where $\partial \boldsymbol{f}_i/\partial \hat{\boldsymbol{\lambda}}$ denotes $\partial \boldsymbol{f}_i/\partial \boldsymbol{\lambda}$ evaluated at $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}$. Consistency of the solution of (5.13), $\hat{\boldsymbol{\lambda}}$, depends upon the correct specification of $E(y_{ij})$ and $E(r_{ij})$.

When $\boldsymbol{r}_i = \{y_{is}y_{it}\}$ and when we use the odds ratio as a measure of association, $\boldsymbol{H}_1$ simplifies to

$$\boldsymbol{H}_1 = \sum_{i=1}^{n} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \boldsymbol{D}_i.$$

5.5 Estimating Equations by a Class of Quadratic Models

Zhao and Prentice (1990) [171] and Prentice and Zhao (1991) [110] introduced a class of quadratic exponential models (see Gourieroux *et al.* (1984) [50]), which were used to develop estimating equations for the mean and covariance parameters. This more systematic approach was compared to the more ad hoc method of GEE for the covariance matrix.

Let $r_{ist} = (y_{is} - \mu_{is})(y_{it} - \mu_{it})$, that is r_{ist} are the empirical covariance values, so that $\boldsymbol{r}_i = \{r_{ist}\}$ for discrete and continuous responses. The mean of $\boldsymbol{r}_i$ is the vector of covariances $\boldsymbol{\sigma}_i = \{\sigma_{ist}\}$ which is a function of parameter vector $\boldsymbol{\alpha}$. The ad hoc method described in Prentice and Zhao (1991) [110] is GEE1 defined in (5.14), where $\boldsymbol{\rho}_i$ is replaced by $\boldsymbol{\sigma}_i$.

Estimating equations by the more systematic approach for the mean and covariance parameters of $\boldsymbol{y}_i$, can be generated by the quadratic exponential model

$$\Pr(\boldsymbol{y}_i; \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i) = \Delta_i^{-1} \exp\left(\boldsymbol{y}_i'\boldsymbol{\theta}_i + \boldsymbol{w}_i'\boldsymbol{\xi}_i + c_i(\boldsymbol{y}_i)\right) \tag{5.16}$$

where $\boldsymbol{w}_i = (y_{i1}^2,\ y_{i1}y_{i2}, \ldots,\ y_{i2}^2,\ y_{i2}y_{i3}, \ldots)'$, $c_i(\cdot)$ is a shape function and $\Delta_i = \Delta_i(\boldsymbol{\theta}_i,\ \boldsymbol{\xi}_i,\ c_i(\cdot))$ is a normalization constant. The 'canonical' parameters $\boldsymbol{\theta}_i = \boldsymbol{\theta}_i(\boldsymbol{\mu}_i,\ \boldsymbol{\sigma}_i) = (\theta_{i1},\ \theta_{i2}, \ldots,\ \theta_{ip_i})'$ and $\boldsymbol{\xi}_i = \boldsymbol{\xi}_i(\boldsymbol{\mu}_i,\ \boldsymbol{\sigma}_i) = (\xi_{i11},\ \xi_{i12}, \ldots, \xi_{i22},\ \xi_{i23}, \ldots)'$ are expressed as functions of the marginal parameters $(\boldsymbol{\mu}_i,\ \boldsymbol{\sigma}_i)$.

Under mild regularity conditions it can be shown that maximum likelihood estimation of the mean and covariance parameters using any member of the family will be

consistent and asymptotically Normal. This family is unique because the consistency holds even if the quadratic exponential model does not hold. Gourieroux *et al.* (1984) [50] called this method pseudo-maximum likelihood estimation.

The Jacobian for the transformation from $(\boldsymbol{\theta}_i,\ \boldsymbol{\xi}_i)$ to $(\boldsymbol{\mu}_i,\ \boldsymbol{\sigma}_i)$ is the inverse of the covariance matrix for $(\boldsymbol{y}_i',\ \boldsymbol{w}_i')'$. This is generally a one-to-one transformation, except if $(\boldsymbol{y}_i',\ \boldsymbol{w}_i')'$ has a degenerate distribution.

Using any particular case of (5.16), the score equations for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are n^{-1} times equations (5.13) with $\boldsymbol{\rho}_i$ replaced by $\boldsymbol{\sigma}_i$, that is they correspond to GEE2. Estimation of the mean and covariance parameters under this pseudo-maximum likelihood approach are, however, typically computationally involved and thus unattractive. A more convenient approach is to replace $\boldsymbol{\Sigma}_i$ by a working covariance matrix as done in GEE2. Prentice and Zhao (1991) [110] offered a number of suggestions for the working covariance matrix $\boldsymbol{V}_i$. They were:

(**WC1**) Independence Working Matrices: Assuming the elements of $\boldsymbol{y}_i$ are independent and using the Normal theory value for r_{ijj} leads to

$$\text{cov}(\boldsymbol{y}_i, \boldsymbol{r}_i) = \boldsymbol{0} \quad \text{and} \quad \text{cov}(r_{ist}, r_{imn}) = 0 \quad s \neq m,\ t \neq n$$

$$\text{var}(r_{ijk}) = \sigma_{ij}\sigma_{ik} \quad j \neq k \quad \text{and} \quad \text{var}(r_{ijj}) = 2\sigma_{ijj}^2.$$

If the dependencies among the elements of $\boldsymbol{y}_i$ are not very great then this specification may be adequate.

(**WC2**) Normal Distribution: Using a Normal distribution for $\boldsymbol{y}_i$ gives

$$\text{cov}(\boldsymbol{y}_i, \boldsymbol{r}_i) = \boldsymbol{0},$$

$$\text{cov}(r_{ist}, r_{imn}) = \text{cov}(y_{is}, y_{im})\text{cov}(y_{it}, y_{in}) + \text{cov}(y_{is}, y_{in})\text{cov}(y_{it}, y_{im}).$$

If the distribution of $\boldsymbol{y}_i$ does not differ greatly from Normality, then we can expect that the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are highly efficient.

(**WC3**) Common Third and Fourth-Order Correlations: This specification is a relaxation of WC2, where

$$E((y_{ij}-\mu_{ij})(y_{ik}-\mu_{ik})(y_{il}-\mu_{il})) = \gamma_{jkl}(\sigma_{ijj}\sigma_{ikk}\sigma_{ill})^{1/2}$$

and

$$E((y_{ij}-\mu_{ij})(y_{ik}-\mu_{ik})(y_{il}-\mu_{il})(y_{im}-\mu_{im})) = \\ \sigma_{ijk}\sigma_{ilm}+\sigma_{ijl}\sigma_{ikm}+\sigma_{ijm}\sigma_{ikl}+\delta_{jklm}(\sigma_{ijj}\sigma_{ikk}\sigma_{ill}\sigma_{imm})^{1/2}$$

where γ_{jkl} and δ_{jklm}, $j \le k \le l \le m$, are the extra parameters that need to be estimated. We can use the $n^{1/2}$-consistent estimators $\hat{\gamma}_{jkl}$ obtained by averaging over all n (such that $p_i \ge l$)

$$(y_{ij}-\mu_{ij})(y_{ik}-\mu_{ik})(y_{il}-\mu_{il})/(\sigma_{ijj}\sigma_{ikk}\sigma_{ill})^{1/2}$$

evaluated at $(\hat{\beta},\ \hat{\gamma})$.

Similarly we can obtain a $n^{1/2}$-consistent estimator of $\hat{\delta}_{jklm}$, where we average over all n such that $p_i \ge m$. This specification allows adaptation to skewness or kurtosis in the sampling distribution, relative to the Normal distribution. Sample sizes (n) may need to be large for good asymptotic efficiencies. Simplifications could be made by making certain of the γ_{jkl} and δ_{jklm} equal, for example if the elements of $\boldsymbol{y}_i$ are in some way exchangeable.

5.6 Efficiency and Consistency of GEE1 and GEE2

The usual focus in longitudinal studies is estimating the mean parameters, and the association parameters are considered as nuisance parameters. By ignoring information on $\boldsymbol{\beta}$ in $\boldsymbol{r}_i$, the use of GEE1 results in consistent estimates of $\boldsymbol{\beta}$ given only that $E(\boldsymbol{y}_i)$ is correctly specified, irrespective of whether $E(\boldsymbol{r}_i)$ is correct or not. The penalty paid by ignoring this information is that these estimators will be less efficient than $\hat{\boldsymbol{\lambda}}$ when

$E(\boldsymbol{r}_i)$ is correct. However the loss of efficiency may be of secondary importance to the inconsistency of the estimate of $\hat{\boldsymbol{\lambda}}$ because of incorrect specification of the expectations.

In a series of simulations of blocked correlated binary data under various models, Liang *et al.* (1992) [82] examined the efficiency of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ relative to maximum likelihood estimates for GEE1 and GEE2, where the odds ratio is a linear function of the parameter vector $\boldsymbol{\alpha}$. They reported estimates of $\boldsymbol{\beta}$ for both methods were highly efficient but there is a non-negligible loss of efficiency in the estimation of $\boldsymbol{\alpha}$ when using separate GEE (GEE1) (as low as 40% when the true odds ratio was 5). This is analogous to using PREML instead of full REML for the profile model discussed in chapter 3. Liang *et al.* (1992) [82] recommend using GEE2 when the number of repeated measurements (p) is large relative to the number of subjects (n) and/or the association parameter vector $\boldsymbol{\alpha}$ is of interest.

Both GEE1 and GEE2 require additional third and fourth moment assumptions. However, GEE1 has the advantage of fewer higher moment assumptions than GEE2. In the commentary of Fitzmaurice, Laird and Rotnitzky (1993) [44], Prentice and Mancl argue that in most situations there is likely to be little usable information of $\boldsymbol{\beta}$ from the vector of empirical covariances. They report that in an extensive set of simulation studies (Mancl (1992) [90]) there was very little efficiency gain in using GEE2 rather than GEE1. In some cases there was a noticeable efficiency loss for estimation of $\boldsymbol{\beta}$ with GEE2 compared to GEE1, even when the variance model was correctly specified. This was apparently due to inappropriate working models for $\boldsymbol{V}_{br}$ and $\boldsymbol{V}_r$ assigning undue weight to the contribution of the $\boldsymbol{r}_i$'s in the estimation of $\boldsymbol{\beta}$. In the rejoinder, Fitzmaurice, Laird and Rotnitzky however express the belief that these findings heavily depend on the choice of design matrix and parameter values.

Another important consideration is the stability of the estimating procedure under GEE1 and GEE2. This will be examined further in chapter 7.

Chapter 6

Extensions of the Generalized Estimating Equation Method

6.1 Introduction

The profile and sum of profile models (chapters 2 and 3) allow for modelling of effects over time for Normal data. In the discrete case, Stram, Wei and Ware (1988) [137] considered the analogous problem, but their analysis is equivalent to GEE with independence 'working' covariance matrix (Zeger (1988) [166]). This will lead to inefficient estimation, much like the case of ordinary least squares for correlated data. The model parameters in Stram, Wei and Ware (1988) [137] were assumed to be occasion-specific, that is they are specific to each occasion.

Thall and Vail (1990) [142] used GEE's to model count data with overdispersion and described some covariance models with occasion-specific variance parameters. Some of the variance models they considered were $\text{var}(y_{it}) = \alpha_t \mu_t$ and $\text{var}(y_{it}) = \alpha_t \mu_t + \alpha_0 \mu^2$, t indexing the repeated measurements within each unit.

In this chapter we will consider the extension of GEE's to models which have occasion-specific mean and/or dispersion parameters. The within unit correlations are not restricted to the independence case. Modelling the occasion-specific mean param-

eters leads to a form analogous to a sum of profiles model and a two-stage estimation procedure may be used. Missing data for equally spaced measurements over time can also be easily accommodated into the analysis.

Modelling the dispersion parameter for univariate data was briefly discussed in §5.2.3, in relation to quasi-likelihood and extended quasi-likelihood. For longitudinal data, a set of GEE's to model the dispersion parameters is developed in this chapter. The modelling of the dispersion parameters may provide important information on the influence of particular covariates on the variance. Another benefit may be an increase in the efficiency in the estimation of the mean parameters. It is noted that a similar set of estimating equations to ours for the dispersion parameters was independently developed by Paik (1992) [104].

A difficulty with the GEE approach is selecting a best model from a number of possible candidate models. Model selection criteria for the GEE will be discussed in this chapter and applied to the examples in chapter 7.

6.2 Occasion-Specific Parameters and Sum of Profiles

Suppose the mean relationship (5.5) is generalized to allow for occasion-specific regression parameters, that is

$$\mu_{ij} = h(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{.j}) \tag{6.1}$$

or equivalently

$$\mu^*_{ij} = f(\mu_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta}_{.j}, \tag{6.2}$$

$\boldsymbol{\beta}_{.j}$ denoting the occasion-specific regression parameters. Stram *et al.* (1988) [137] developed an approach that is closely related to that of Liang and Zeger (1986) [81]. Firstly they analysed the data at each occasion separately using a proportional-odds model from the family of models for ordinal data described by McCullagh (1980) [91]. They combined the estimates of the occasion-specific parameters $\boldsymbol{\beta}_{.j}$ into a single vector $\hat{\boldsymbol{\beta}}_c$

say, which is asymptotically unbiased and Normal, and performed simultaneous inferences on individual characteristics over time. Zeger (1988) pointed out that by allowing the regression parameters to vary in time and using the working correlation matrix, $\boldsymbol{R}(\boldsymbol{\alpha}) = \boldsymbol{I}$, then $\hat{\boldsymbol{\beta}}_c$ is the solution of (5.8).

We formulate the GEE for the occasion-specific problem in the following manner. Suppose the ith individual is measured at p_i of a total of p time points. Let $\boldsymbol{\mu}_i^* = (\mu_{i1}^*, \ldots, \mu_{ip_i}^*)'$ and $\boldsymbol{B} = (\boldsymbol{\beta}_{.1}, \ldots, \boldsymbol{\beta}_{.p}) = (\boldsymbol{\beta}_{1.}, \ldots, \boldsymbol{\beta}_{m.})'$. For the ith individual we can write the linear forms in (6.2) as $\boldsymbol{\mu}_i^* = \boldsymbol{E}_i\boldsymbol{\beta}$, where $\boldsymbol{E}_i$ is a $p_i \times mp$ design matrix that reflects both the treatments applied and the time points at which observations have been made. This formulation allows missing values to be easily accommodated. The $mp \times 1$ vector $\boldsymbol{\beta}$ is the vector of parameters for the time-by-treatment combinations obtained by stacking the rows of $\boldsymbol{B}$, that is $\boldsymbol{\beta} = (\boldsymbol{\beta}_{1.}', \ldots, \boldsymbol{\beta}_{m.}')'$. As a consequence we may write (6.1) as

$$\mu_{ij} = h(\boldsymbol{e}_{ij.}'\boldsymbol{\beta}),$$

where $\boldsymbol{e}_{ij.}'$ is the jth row of $\boldsymbol{E}_i$. We now have exactly the same form as that in §5.3.2 and hence the results follow here, except the divisors for $\hat{\phi}$ and $\hat{R}_{st}$ are now $(n-m)p$ and $n - mp$ respectively. We could also generalize (6.1) (and similarly (6.2)) to $\mu_{ij} = h_j(\boldsymbol{x}_{ij}'\boldsymbol{\beta}_{.j})$ as discussed by Liang *et al.* (1992) [82].

To investigate differences over time for intercept terms of the occasion-specific mean model, it may be better to use a *deviations from the sample averages* (DFTSA) form (Weisberg (1980) [160] (pp. 10-11)) of the mean model. For a simple linear model,

$$y_i = \beta_0 + \beta_1 x_i + e_i,$$

the DFTSA form is

$$y_i = \beta_0^* + \beta_1(x_i - \bar{x}) + e_i,$$

where e_i is the random error term, and $\beta_0^* = \beta_0 + \beta_1\bar{x}$. Thus the new occasion-specific intercept terms are now taken about the covariate sample averages (reducing the effect of time-varying parameters of the covariates on the intercept estimates).

Now suppose that we are interested in modelling the blocks of $\boldsymbol{\beta}$ by a linear model of the form (2.23), that is

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_{1.} \\ \boldsymbol{\beta}_{2.} \\ \vdots \\ \boldsymbol{\beta}_{m.} \end{pmatrix} = \begin{pmatrix} \boldsymbol{M}_1\boldsymbol{\theta}_1 \\ \boldsymbol{M}_2\boldsymbol{\theta}_2 \\ \vdots \\ \boldsymbol{M}_m\boldsymbol{\theta}_m \end{pmatrix} = \begin{pmatrix} \boldsymbol{M}_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{M}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{M}_m \end{pmatrix} \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \\ \vdots \\ \boldsymbol{\theta}_m \end{pmatrix} = \boldsymbol{M}\boldsymbol{\theta} \qquad (6.3)$$

and so the profile design matrices $\boldsymbol{M}_i^{p\times q_i}$ allow different parametric forms for different blocks of $\boldsymbol{\beta}$. This allows for modelling of individual characteristics over time. We can write (6.2) in a form analogous to the sum of profiles model (2.24),

$$\mu_{ij}^{*} = \sum_{l=1}^{m} x_{ijl}\boldsymbol{\theta}_l'\boldsymbol{m}_{lj.}$$

where $\boldsymbol{m}_{lj.}$ is the jth row of $\boldsymbol{M}_l$.

The GEE when (6.3) is true is

$$\sum_{i=1}^{n} \boldsymbol{U}_{ti}(\boldsymbol{\theta},\ \boldsymbol{\alpha}) = \sum_{i=1}^{n} (\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\theta})'\boldsymbol{V}_{bi}^{-1}\boldsymbol{S}_{bi} = \boldsymbol{0}. \qquad (6.4)$$

With time-independent explanatory variables, Gaussian responses and identity link function, this model is a form of (2.24) and (6.4) is the score function of the sum of profiles model. Computation of $\hat{\boldsymbol{\theta}}_R$ is as in §5.3.2 for $\hat{\boldsymbol{\beta}}_R$.

The estimation procedure for $\boldsymbol{\theta}$ from the GEE's is two-stage. First we estimate $\boldsymbol{\beta}$ via the GEE (5.7) to obtain the estimator $\hat{\boldsymbol{\beta}}_R$. We then test and select the appropriate profile specification in (6.3) and then use (6.4) to obtain the estimator $\hat{\boldsymbol{\theta}}_R$ of $\boldsymbol{\theta}$. Using the asymptotic distribution of $\hat{\boldsymbol{\beta}}_R$ we can perform a goodness of fit test of (6.3), that is

$$(\hat{\boldsymbol{\beta}}_R - \boldsymbol{M}\boldsymbol{\theta})'\boldsymbol{V}_R^{-1}(\hat{\boldsymbol{\beta}}_R - \boldsymbol{M}\boldsymbol{\theta})$$

which has an asymptotic χ^2 distribution on $mp-q$ degrees of freedom, where $q = \sum_{i=1}^{m} q_i$. We can also perform tests on hypotheses of the form $\boldsymbol{C}\hat{\boldsymbol{\beta}}_R = \boldsymbol{0}$ where $\boldsymbol{C}$ is a known $r \times mp$ matrix, using the test statistic

$$(\boldsymbol{C}\hat{\boldsymbol{\beta}}_R)'(\boldsymbol{C}\boldsymbol{V}_R\boldsymbol{C}')^{-1}\boldsymbol{C}\hat{\boldsymbol{\beta}}_R \qquad (6.5)$$

which has an asymptotic χ^2 distribution on r degrees of freedom.

If the number of regression parameters and/or time points are moderately large, then the selection of the appropriate profile specification could be performed in stages, analogous to the standard backwards elimination method. For example the assumption of separate regression parameters at each occasion could be tested in stages by applying (6.5) to each characteristic (adjusting the contrast matrix $\boldsymbol{C}$ for each characteristic). The smallest non-significant test statistic indicates an intermediate profile specification.

At this stage one of two different strategies could be employed. Under the first strategy new estimates are obtained by the GEE (6.4) and the variance matrices are adjusted or updated accordingly. New tests based on (6.5) ($\hat{\boldsymbol{\theta}}_R$ replacing $\hat{\boldsymbol{\beta}}_R$) are performed on $\boldsymbol{\theta}$. This procedure is continued until the final profile model specification is reached.

Alternatively, a computationally simpler strategy is to use the asymptotic properties of the GEE estimate $\hat{\boldsymbol{\beta}}_R$ to estimate $\boldsymbol{\theta}$ by generalized least squares. If (6.3) is correct, then $\hat{\boldsymbol{\beta}}_R$ is asymptotically distributed as $N(\boldsymbol{M\theta},\ \boldsymbol{V}_R)$. The generalized least squares estimator of $\boldsymbol{\theta}$ based on the asymptotic distribution of $\hat{\boldsymbol{\beta}}_R$ is

$$\hat{\boldsymbol{\theta}}^* = (\boldsymbol{M}'\boldsymbol{V}_R^{-1}\boldsymbol{M})^{-1}\boldsymbol{M}'\boldsymbol{V}_R^{-1}\hat{\boldsymbol{\beta}}_R$$

and $\hat{\boldsymbol{\theta}}^*$ has asymptotic covariance matrix $(\boldsymbol{M}'\boldsymbol{V}_R^{-1}\boldsymbol{M})^{-1}$. This is analogous to the generalized least squares estimator (2.25). New tests on $\hat{\boldsymbol{\theta}}^*$ are implemented and a new estimate of $\boldsymbol{\theta}$ is obtained by generalized least squares. This procedure is repeated until the final profile model specification is determined. The main difference between this strategy and the previous one is that the estimating equations are employed once only and consequently $\boldsymbol{V}_R$ is evaluated once.

The second strategy is probably the less efficient of the two. However it would be more robust to an incorrect specification of the profile design matrix $\boldsymbol{M}$. In the first strategy, the association and dispersion parameter estimates are re-estimated at each step. If the profile design specification is incorrect at one of the stages then these new parameter estimates are now incorrect, and inferences on the mean parameters may be misleading. This is in contrast to the second strategy, where the association and

dispersion parameters are estimated once only.

If we are interested in modelling the characteristics ($\boldsymbol{\beta}_{.k}$, $k = 1, \ldots, p$) at each time point, we can model the $pm \times 1$ vector $\boldsymbol{\xi}$ attained by stacking the columns of $\boldsymbol{B}$, which is just a reordering of $\boldsymbol{\beta}$, in a similar fashion as (6.3) for $\boldsymbol{\beta}$. The estimator $\hat{\boldsymbol{\xi}}$ of $\boldsymbol{\xi}$ derived from the GEE's (5.8) is also a reordering of the elements of $\hat{\boldsymbol{\beta}}_R$.

6.3 A Dispersion Model

Suppose that we replace the dispersion parameter ϕ in (5.6) by ϕ_{ij}. Generalizing the univariate case of §5.2.3, let $d_{ijj} = (y_{ij} - \mu_{ij})^2 / g(\mu_{ij})$, which has expectation ϕ_{ij}. Assume that we can model the dispersion parameters ϕ_{ij} by

$$b(\phi_{ij}) = \boldsymbol{a}_{ij}' \boldsymbol{\gamma} \tag{6.6}$$

where b is a link function, $\boldsymbol{a}_{ij}$ is a $k \times 1$ vector of covariates and $\boldsymbol{\gamma}$ is the corresponding vector of unknown parameters. Let $\boldsymbol{\phi}_i = (\phi_{i1}, \cdots, \phi_{in_i})'$ and $\boldsymbol{d}_i = (d_{i11}, \cdots, d_{in_in_i})'$. To complete our second-order construction, we use the empirical pairwise correlation values $r_{ist} = (y_{is} - \mu_{is})(y_{it} - \mu_{it}) / (\phi_{is}\phi_{it} g(\mu_{is}) g(\mu_{it}))^{1/2}$ where $s \neq t$, for the response vector $\boldsymbol{r}_i = \{r_{ist}\}$. We assume that $\rho_{ijk} = c(\boldsymbol{\alpha})$, where ρ_{ijk} is the correlation between y_{ij} and y_{ik}. The unknown $s \times 1$ parameter vector $\boldsymbol{\alpha}$ specifies the pairwise correlations and the inverse of c is a link function. Since only first and second order moment assumptions have been made, the covariance matrix of $\boldsymbol{d}_i$ and $\boldsymbol{r}_i$ is unknown, involving fourth order moments. However we can use working covariance matrices $\boldsymbol{V}_{di}$ and $\boldsymbol{V}_{ri}$ for $\boldsymbol{d}_i$ and $\boldsymbol{r}_i$ respectively.

The dispersion regression parameters $\boldsymbol{\gamma}$ can be estimated from the GEE

$$\sum_{i=1}^{n} \boldsymbol{U}_{di}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\phi}_i}{\partial \boldsymbol{\gamma}} \right)' \boldsymbol{V}_{di}^{-1} (\boldsymbol{d}_i - \boldsymbol{\phi}_i) = \boldsymbol{0} \tag{6.7}$$

and the correlation specification parameter vector $\boldsymbol{\alpha}$ from (5.11), where $\boldsymbol{\rho}_i = \{\rho_{ist}\}$. It may be the case that the dispersion regression parameters are occasion-specific, that is

$$b(\phi_{ij}) = \boldsymbol{a}_{ij}' \boldsymbol{\gamma}_{.j},$$

which allows an extension to the sum of profiles procedure described in §6.2 for the $\gamma_{.j}$'s.

Equation (6.7) is of the same form as that given independently in Paik (1992) [104]. The mean regression parameters are estimated by GEE (5.7) and the correlation parameters are estimated by the method of moments. The actual estimating procedure used is GEE1.

If we let $\boldsymbol{\lambda} = (\boldsymbol{\beta}',\ \boldsymbol{\gamma}',\ \boldsymbol{\alpha}')'$ and $\boldsymbol{f}_i = (\boldsymbol{y}_i',\ \boldsymbol{d}_i',\ \boldsymbol{r}_i')'$ with mean vector $\boldsymbol{\eta}_i = (\boldsymbol{\mu}_i',\ \boldsymbol{\phi}_i',\ \boldsymbol{\rho}_i')'$ then we can use (5.7), (6.7) and (5.11) to estimate $\boldsymbol{\lambda}$, that is in the spirit of GEE1. Note that the mean parameter vector $\boldsymbol{\beta}$ may refer to occasion-specific parameters. The parameter vector $\boldsymbol{\lambda}$ may also be estimated from the unbiased estimating equations (5.13), where

$$\boldsymbol{D}_i = \partial\boldsymbol{\eta}_i/\partial\boldsymbol{\lambda} = \begin{bmatrix} \boldsymbol{D}_{bi} & 0 & 0 \\ \partial\boldsymbol{\phi}_i/\partial\boldsymbol{\beta} & \boldsymbol{D}_{di} & 0 \\ \partial\boldsymbol{\rho}_i/\partial\boldsymbol{\beta} & \partial\boldsymbol{\rho}_i/\partial\boldsymbol{\gamma} & \boldsymbol{D}_{ri} \end{bmatrix},$$

$\boldsymbol{D}_{di} = \partial\boldsymbol{\phi}_i/\partial\boldsymbol{\gamma}$. This is an extension of GEE2. In the examples in chapter 7, all models considered result in $\partial\boldsymbol{\rho}_i/\partial\boldsymbol{\beta} = \boldsymbol{0}$ and $\partial\boldsymbol{\rho}_i/\partial\boldsymbol{\gamma} = \boldsymbol{0}$.

Replace $\boldsymbol{\Sigma}_i$ in (5.13) by the 'working' covariance matrix

$$\boldsymbol{V}_i = \begin{bmatrix} \boldsymbol{V}_{bi} & \boldsymbol{V}_{bdi} & \boldsymbol{V}_{bri} \\ \cdot & \boldsymbol{V}_{di} & \boldsymbol{V}_{dri} \\ \cdot & \cdot & \boldsymbol{V}_{ri} \end{bmatrix},$$

and let the $v \times 1$ parameter vector $\boldsymbol{\xi}$ completely specify the matrices $\boldsymbol{V}_{bdi}$, $\boldsymbol{V}_{bri}$, $\boldsymbol{V}_{dri}$, $\boldsymbol{V}_{di}$ and $\boldsymbol{V}_{ri}$. Let $\hat{\boldsymbol{\xi}}(\boldsymbol{\lambda})$ be an $n^{1/2}$-consistent estimate $\boldsymbol{\xi}$, and consequently let $\hat{\boldsymbol{\lambda}}$ be the solution of (5.13).

The following definition is used in the proof of Theorem 6.1.

Definition 6.1 *For a sequence of random variables $\{a_n\}$, then $o_p(a_n)$ represents a random variable such that for any fixed $\epsilon > 0$,*

$$P(|o_p(a_n)/a_n| \leq \epsilon) \to 1 \text{ as } n \to \infty,$$

that is $o_p(a_n)/a_n$ converges in probability to 0.

A simple modification of Theorem 5.1 in §5.3.2 to account for the second order (dispersion and correlation) extensions under GEE2 is formally presented in the following theorem. The proof is a straightforward extension of the proof of Theorem 5.1 (Liang and Zeger (1986) [81]).

Theorem 6.1 *Under mild regularity conditions, and* $\hat{\boldsymbol{\xi}}$ *is* $n^{1/2}$*-consistent given* $\boldsymbol{\lambda}$*, then the vector* $n^{1/2}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda})$ *is asymptotically multivariate Normal with zero mean and covariance matrix given by*

$$\boldsymbol{\Omega} = \lim_{n\to\infty} n(\boldsymbol{\Omega}_1^{-1}\boldsymbol{\Omega}_0\boldsymbol{\Omega}_1^{-1}), \tag{6.8}$$

where

$$\boldsymbol{\Omega}_0 = \sum_{i=1}^{n} \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\mathrm{cov}(\boldsymbol{f}_i)\boldsymbol{V}_i^{-1}\boldsymbol{D}_i \quad \textit{and} \quad \boldsymbol{\Omega}_1 = \sum_{i=1}^{n} \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}(\boldsymbol{D}_i + \partial\boldsymbol{f}_i/\partial\boldsymbol{\lambda}).$$

Proof: An outline of the proof is given in Appendix A.

The asymptotic covariance matrix (6.8) can be consistently estimated by (5.15) if the mean vector is correctly specified. Consistency of the solution of (5.13), $\hat{\boldsymbol{\lambda}}$, depends upon the correct specification of $E(y_{ij})$, $E(d_{ij})$ and $E(r_{ij})$ and as such may not be robust to incorrect specification.

The matrix

$$\partial\boldsymbol{f}_i/\partial\boldsymbol{\lambda} = \begin{bmatrix} 0 & 0 & 0 \\ \partial\boldsymbol{d}_i/\partial\boldsymbol{\beta} & 0 & 0 \\ \partial\boldsymbol{r}_i/\partial\boldsymbol{\beta} & \partial\boldsymbol{r}_i/\partial\boldsymbol{\gamma} & 0 \end{bmatrix}$$

from Appendix A may be simplified further. Since

$$\frac{\partial d_{ij}}{\partial\beta_k} = -2\frac{(y_{ij}-\mu_{ij})}{g(\mu_{ij})}\frac{\partial\mu_{ij}}{\partial\beta_k} - \frac{(y_{ij}-\mu_{ij})^2}{g(\mu_{ij})^2}\frac{\partial g(\mu_{ij})}{\partial\beta_k} \tag{6.9}$$

then part of $\sum_{i=1}^{n} \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\partial\boldsymbol{f}_i/\partial\boldsymbol{\lambda}$ corresponds to sample averages of linear functions of the $(y_{ij}-\mu_{ij})$ (from the first term of (6.9)) and is $o_p(1)$. Thus for large sample sizes, we can effectively set the first term in (6.9) to zero. A similar simplification may be used in $\partial\boldsymbol{r}_i/\partial\boldsymbol{\beta}$. If instead the d_{ij} are the empirical variances $(y_{ij}-\mu_{ij})^2$, then by the same arguments $\partial\boldsymbol{d}_i/\partial\boldsymbol{\beta}$ is approximately zero. Similarly if $r_{ijk} = (y_{ij}-\mu_{ij})(y_{ik}-\mu_{ik})$ then $\partial\boldsymbol{r}_i/\partial\boldsymbol{\beta}$ is approximately zero and $\partial\boldsymbol{r}_i/\partial\boldsymbol{\gamma}$ is exactly zero.

Using GEE1 type estimation for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ instead of GEE2 (5.13) is equivalent to presuming $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ are mutually orthogonal. As discussed in §5.6 there is a trade-off to be considered between the increased efficiency under GEE2, and the robustness under GEE1.

The use of (5.7), (6.7) and (5.11) is equivalent to requiring the following restrictions,

$$\partial\boldsymbol{\phi}_i/\partial\boldsymbol{\beta} = \mathbf{0}, \quad \partial\boldsymbol{\rho}_i/\partial\boldsymbol{\beta} = \mathbf{0} \quad \text{and} \quad \partial\boldsymbol{\rho}_i/\partial\boldsymbol{\gamma} = \mathbf{0}, \tag{6.10}$$

and

$$\boldsymbol{V}_{bdi} = \mathbf{0}, \quad \boldsymbol{V}_{bri} = \mathbf{0} \quad \text{and } \boldsymbol{V}_{dri} = \mathbf{0}. \tag{6.11}$$

If $\tilde{\boldsymbol{\lambda}}$ is the solution of (5.7), (6.7) and (5.11), then by Theorem 6.1, $n^{1/2}(\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda})$ is asymptotically Normal with zero mean and variance (6.8), under the restrictions (6.10) and (6.11). Depending on the need, efficiency may be increased and robustness decreased or vice versa by using separate GEE's (GEE1) or a combined GEE (GEE2), or a mixture of both such as estimating $\boldsymbol{\beta}$ by (5.7) and a combined GEE that estimates $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$. In this case the estimate of $\boldsymbol{\beta}$ is consistent irrespective of whether the correct mean specifications for $\boldsymbol{\phi}_i$ and $\boldsymbol{\rho}_i$ are given. Other GEE combinations may be useful for different levels of parameter importance. Estimating the parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ in a combined GEE separate from $\boldsymbol{\beta}$ is analogous to estimating the dispersion and correlation parameters by REML in Normal Theory, that is from the marginal likelihood of zero contrasts. Similarly, the mean and dispersion parameters are estimated by separate quasi-likelihoods in Smyth (1989) [131].

In §5.5 we listed a number of possible working covariance matrices (WC1, WC2 and WC3) as suggested by Prentice and Zhao (1991) [110] for the vector of empirical covariances, and so leading to working covariance matrices $\boldsymbol{V}_{di}$ and $\boldsymbol{V}_{ri}$. Another possibility for $\boldsymbol{V}_{di}$ would be to assume that the correlation matrix of $\boldsymbol{d}_i$ follows an autoregressive structure of order 1 (Paik (1992) [104]). Unfortunately a sensible correlation structure for $\boldsymbol{r}_i$ analogous to AR-1 is not so simple to define but the approximations WC2 or WC3 may still be adequate for $\boldsymbol{V}_{ri}$ (and $\boldsymbol{V}_{bri}$ if GEE2 is used). Other working covariance matrices of $\boldsymbol{f}_i$ could also be used (see Prentice and Zhao (1991) [110]).

6.4 Pseudo-likelihood Estimates

Suppose the variance function belongs to a family of functions indexed by an unknown parameter ζ as in the univariate case of §5.2.2. For longitudinal data, then the variance can be written as

$$\text{var}(y_{ij}) = \phi_{ij} g_\zeta(\mu_{ij}),$$

where for example

$$g_\zeta(\mu_{ij}) = \mu_{ij}^\zeta$$

where common values of ζ are 0, 1, 2 and 3. To estimate ζ, we can consider the pseudo-likelihood approach of Davidian and Carroll (1987) [31]. If we condition on the GEE estimates $\hat{\boldsymbol{\lambda}}$ and use the 'working covariance' matrix for the responses f_{ij} then assuming a multivariate Normal distribution allows us to find an estimate of ζ, in a straightforward manner. Further, by conditioning only on the GEE estimate $\hat{\boldsymbol{\beta}}$ via (5.7), we can find pseudo-likelihood estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ as well (see Thall and Vail (1990) [142]). This offers an alternative approach to estimating the nuisance parameters but will not be considered further.

6.5 Estimation of the GEE parameters

If $\hat{\boldsymbol{\lambda}}_0 = (\hat{\boldsymbol{\beta}}_0,\ \hat{\boldsymbol{\gamma}}_0,\ \hat{\boldsymbol{\alpha}}_0)'$ are the initial starting values of $\boldsymbol{\lambda} = (\boldsymbol{\beta},\ \boldsymbol{\gamma},\ \boldsymbol{\alpha})$, then $\boldsymbol{\lambda}$ may be estimated iteratively by the modified scoring procedure (j^{th} iteration),

$$\begin{aligned}
\hat{\boldsymbol{\lambda}}_j &= \hat{\boldsymbol{\lambda}}_{j-1} + \Big(\sum_{i=1}^{n} \partial U_i(\boldsymbol{\lambda}_{j-1},\ \hat{\boldsymbol{\xi}}(\boldsymbol{\lambda}_{j-1}))/\partial \boldsymbol{\lambda}_{j-1}\Big)^{-1} \sum_{i=1}^{n} U_i(\boldsymbol{\lambda}_{j-1},\ \hat{\boldsymbol{\xi}}(\boldsymbol{\lambda}_{j-1})) \\
&= \hat{\boldsymbol{\lambda}}_{j-1} + \Big(\sum_{i=1}^{n} \boldsymbol{D}'_{i(j-1)} \boldsymbol{V}^{-1}_{i(j-1)} (\boldsymbol{D}_{i(j-1)} - \partial \boldsymbol{f}_i/\partial \boldsymbol{\lambda}_{j-1})\Big)^{-1} \\
&\quad \times \sum_{i=1}^{n} U_i(\boldsymbol{\lambda}_{j-1},\ \hat{\boldsymbol{\xi}}(\boldsymbol{\lambda}_{j-1})), \qquad (6.12)
\end{aligned}$$

where the product terms on the right-hand side of the equal sign are evaluated at the $(j-1)^{th}$ value, $\hat{\boldsymbol{\lambda}}_{j-1}$. Instead of (6.12), setting $\partial \boldsymbol{f}_i/\partial \boldsymbol{\lambda} = \boldsymbol{0}$ generally leads to a procedure

both faster and more stable in terms of convergence, especially for small data sets. This modification is also computationally simpler.

6.6 Moment Estimation

The moment estimators of the correlation parameters are derived as follows: suppose that s_{ij} is the standardized residual,

$$s_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\hat{\phi}_{ij}^{1/2} g(\hat{\mu}_{ij})^{1/2}}.$$

When $p_i = p \ \forall \ i$, and $\boldsymbol{R}(\boldsymbol{\alpha}) = \boldsymbol{R}_i(\boldsymbol{\alpha})$ is the unstructured or unspecified correlation matrix, then the $p(p-1)/2$ unique correlation parameters may be estimated by the moment estimator

$$\hat{R}_{jk} = \frac{\sum_{i=1}^{n} s_{ij} s_{ik}}{(\sum_{i=1}^{n} s_{ij}^2 \sum_{i=1}^{n} s_{ik}^2)^{1/2}}, \tag{6.13}$$

given $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. Liang and Zeger (1986) [81], suggested using the moment estimator

$$\hat{\boldsymbol{R}}(\boldsymbol{\alpha}) = \frac{1}{n\hat{\phi}} \sum_{i=1}^{n} \boldsymbol{A}^{-1/2} \boldsymbol{S}_i' \boldsymbol{S}_i \boldsymbol{A}^{-1/2}$$

for scalar ϕ, and which is easily modified for vector $\boldsymbol{\phi}$. This estimator however may give diagonal values unequal to 1. Paik (1992) [104] also uses the moment estimator (6.13), but appears to have incorrectly defined s_{ij} as the square of the above term.

In the case of an equal correlation, $\text{cov}(y_{ij}, y_{ik}) = \alpha \ \forall \ i, \ j$ and k, then the moment estimator of α is

$$\hat{\alpha} = \frac{\sum_{j>k} \hat{R}_{jk}}{p(p-1)/2}.$$

If $\boldsymbol{R}_i(\boldsymbol{\alpha})$ follows an AR-1 structure, that is $\text{cov}(y_{ij}, y_{ik}) = \alpha^{|j-k|}$, then regressing $\log \hat{R}_{jk}$ on $|j-k|$ leads to an estimate of $\log \alpha$, and hence α, from the slope. For more complicated structures, moment estimation can be messy; see for example chapter 8.

6.7 Model Selection Criteria

The statistic $\log(G)$, where $G = \det(\mathrm{var}(\hat{\boldsymbol{\lambda}})^{-1})$ (see Thall and Vail (1990) [142]) can be used to ordinally evaluate goodness of fit, that is higher values correspond to better fits of the data. This is because $\det(\mathrm{var}(\hat{\boldsymbol{\lambda}})^{-1})$ is an increasing function of the squared vector correlation between the estimation equations and the score vector of the true likelihood (which is unknown, as is the squared correlation vector). Thus a higher value of the statistic corresponds to a higher squared vector correlation.

Another way of interpreting the $\log(G)$ statistic is that higher values correspond to more optimal parameter estimation with respect to the variance of the estimators (see Godambe and Heyde (1987) [49]). It can be thought of as a function of the precision of the estimators.

There is however a difficulty in attempting to use $\log(G)$ to compare the various models. As the number of parameters increases, $\log(G)$ also increases. Comparing a model with large numbers of parameters to one with a much smaller set may be deceptive. This is also a problem with maximum likelihood or maximizing some other goodness of fit criteria, as these methods are liable to automatically select the model with highest possible dimension. Some form of parameter size compensation or penalty factor appears to be in order.

A number of methods have been proposed for selecting a model from a set of alternative statistical models, particularly where the alternative models contain different numbers of unknown estimable parameters (different dimensionality). The Kullback-Leibler (K-L) information quantity (Kullback and Leibler (1951) [70])

$$I(f_T;\ f_M) = E\left(\log\left(\frac{f_T(X)}{f_M(X)}\right)\right) = \int_{-\infty}^{\infty} \log\left(\frac{f_T(x)}{f_M(x)}\right) f_T(x)dx,$$

is a measure (metric) of the distance between a model density function and the true underlying density function, where $f_T(x)$ is the true density, and $f_M(x)$ is the density function that defines the model. The negative of the K-L information quantity is called the generalized entropy (Boltzmann (1877) [11]). A model selection criterion is to select

a model density that has the smallest distance from the true unknown density.

The well known Akaike Information Criterion (AIC) was developed by Akaike (1973) [3] as an estimate of the entropy. The AIC can be used to test nested and non-nested models and an allowance is made for parsimony. Let k be the number of estimable parameters for a given model. AIC is defined here as,

Definition 6.2 *AIC: Select the model for which the quantity,*

$$log(maximized\ likelihood) - k,$$

is the largest.

A criticism of the AIC is that it is not consistent. For example, suppose an autoregressive process has order p and the number of time series observations tends to infinity, then frequently AIC does not select the proper order p.

Schwarz (1978) [126] presented an alternative criterion which is known as the Schwarz Criterion (SC) or Bayesian Information Criterion. The justification for this criterion is essentially Bayesian, since asymptotically, we select the model with the largest posterior probability distribution corresponding to a special prior distribution. The a priori distribution is the weighted sum of conditional a priori distributions. The weights are the a priori probabilities that the j^{th} competing model is the true one, and the conditional a priori distributions are the distributions of the random parameter vector given the j^{th} model. See Schwarz (1978) [126] for further details.

Definition 6.3 *SC: Select the model for which the quantity,*

$$log(maximized\ likelihood) - \frac{1}{2}k \log n,$$

is the largest.

This is basically a modified AIC, with increased penalty of overfitting. Kashyap (1982) [62] also takes a Bayesian approach but adds extra terms to SC by going a term further than Schwarz (1978) [126] in the asymptotic expansion of the logarithm of the posterior probabilities.

A slight variation to SC is Bozdogan's (1987) [14] so called consistent AIC (CAIC).

Definition 6.4 *CAIC: Select the model for which the quantity,*

$$\textit{log(maximized likelihood)} - \frac{1}{2}k(1 + \log n),$$

is the largest.

This is derived as an analytical extension to the AIC, without violating the principles laid down by Akaike (1973) [3]. Bozdogan (1987) [14] exploits the large sample asymptotic distributional properties of the maximum likelihood estimators to propose a different information criterion that incorporates the Fisher information matrix into the penalty component of the criterion. It is referred to as the Consistent AIC with Fisher Information (CAICF). Both of these criterions make AIC asymptotically consistent.

CAIC and CAICF should be used if it is desired to avoid overfitting a model. This may be at the expense of underfitting the model sometimes in finite samples, especially for the CAICF. However, because of the consistency, the probability of underfitting and overfitting a model will diminish as the sample size increases. If it is important to avoid underfitting, then use AIC or perhaps CAIC when the sample size is small. The CAICF will not be considered further.

Jones (1993) [61] suggests using AIC, but recommends that models within two units of the highest AIC be taken as competitive models for the best. Among these competitive models, selection is usually based on the model with the smallest number of estimable parameters, k. There is some theoretical justification for this interpretation of AIC (Duong (1984) [34]).

Since no likelihood function is defined with the GEE's, a multivariate Normal distribution will be used with the above information criteria. The advantage in using a multivariate Normal approximation is that it has a simple form, permits the correlations within each unit to be easily included into the information criteria, and in many cases provides a reasonable approximation to non-Normal data.

Chapter 7

Applications of Generalized Estimating Equation Methods

7.1 Introduction

In this chapter we consider three examples. The first example involves a clinical trial. Fifty nine epileptics were observed at four time points, the response being the number of seizures in each time interval. Occasion-specific parameters for the mean and dispersion models will be considered for this example. The second example involves a study of prolactin levels in 30 women, the prolactin levels taken at four time points. The third example consists of counts of the ventricular premature heartbeats for patients that have experienced a myocardial infarction.

Some of the issues considered will be:

(i) The suitability of the different model selection criteria in measuring the success of a model.

(ii) Stability of the GEE estimating procedure, including alternative dispersion/variance model specifications. Related to the issue of stability is the orthogonality of GEE1 compared to GEE2.

(iii) The effect of the choice of the specification of the dispersion/variance model on inferences on the mean parameters. Alternative specifications of the dispersion parameter, ϕ_i, and the variance function, $g(\mu_i)$, were discussed by Firth (1991) [42] (pp. 76-77) for the extended quasi-likelihood Q^+. The alternative specifications can lead to different values for Q^+.

7.2 Examples

7.2.1 Seizure example

Thall and Vail (1990) [142] present analyses of data given in Table 2 of their paper for a clinical trial of 59 epileptics (Leppik *et al.* (1985) [80]). Patients were randomized to receive either an antiepileptic drug, progabide, or a placebo. There were four successive clinical visits after randomization where the number of seizures occurring over the previous 2 weeks since the last visit (or the randomization starting point for the first visit) were recorded. Table 3 in Thall and Vail (1990) [142] shows that the seizure counts display a large amount of extra-Poisson variation (overdispersion), heteroscedasticity and within-patient correlation.

The covariates considered in their models are the baseline seizure rate calculated as the log(psc/4), where psc is the 8-week prerandomization seizure count, log(Age), the binary indicator Trt for the progabide drug group, and the binary indicator Visit4 for the fourth clinical visit.

Thall and Vail (1990) [142] demonstrate a heuristic derivation of a parametric form for the block-diagonal variance matrix $\boldsymbol{V}$ using random effects acting multiplicatively on the mean. Other forms of $\boldsymbol{V}$ are suggested by this derivation.

We consider a number of models and employ GEE's to obtain estimates of the unknown parameters. All the models considered will have the mean form

$$\log(\mu_{ij}) = \beta_{0j} + \beta_{1j}\text{Base} + \beta_{2j}\text{Trt} + \beta_{3j}\text{Base.Trt} + \beta_{4j}\text{Age}, \tag{7.1}$$

that is the parameters are occasion specific. Note that Visit4 cannot be included in the mean because of aliasing with the occasion specific intercept terms (β_{0j}). Mean model (7.1) is the occasion-specific generalization of the mean model (except for Visit4) considered in Thall and Vail (1990) [142]. Let ρ_{ijk} be the correlation between y_{ij} and y_{ik}. For ease of interpretation we will use a DFTSA form (§6.2) of model (7.1) instead, where β^*_{0j} represents the new intercept parameter, all other parameters remaining the same.

The models considered are displayed in Table 7.1, also indicating which working covariance structure, WC1 or WC2 (see §5.5), is used. Note that using WC1 or WC2 implies that the mean parameters are estimated directly from (5.8). This is analogous to REML in Normal theory, where the mean regression parameters are estimated from the conditional likelihood alone, and the dispersion and correlation parameters are estimated from the marginal likelihood where an adjustment over the full likelihood is made. A DFTSA form of dispersion Model 2 (ϵ^*_{0j} replaces ϵ_{0j}) will be used.

Models 1A and 3A are the same as Models 11 and 23 in Table 1 of Thall and Vail (1990) [142]. The variance parameters in Model 3A are estimated by using a variance response vector with elements $(y_{ij} - \mu_{ij})^2$. Equivalently the variance under Model 3A can be rewritten as $\phi_{ij} g(\mu_{ij})$, such that

$$\phi_{ij} = 1 + \phi_j \mu_{ij} \quad \text{and} \quad g(\mu_{ij}) = \mu_{ij}.$$

and is referred to as Model 4A. The dispersion response vector with elements $(y_{ij} - \mu_{ij})^2 / g(\mu_{ij})$ is used for estimation of the variance parameters under Model 4A (Reparameterized Quadratic Model).

The parameters of Models 3A and 4A were estimated by separate GEE (GEE1), that is the mean parameters are estimated from (5.7), and the correlation and variance parameters are estimated in a combined GEE like (5.13). This is the same as setting both $\partial \boldsymbol{\phi}_i / \partial \boldsymbol{\beta}$ and $\partial \boldsymbol{\rho}_i / \partial \boldsymbol{\beta}$ to zero in $\boldsymbol{D}_i$ ($\boldsymbol{V}_{bdi}$ and $\boldsymbol{V}_{bri}$ in $\boldsymbol{V}_i$ are already zero under WC1 or WC2). The parameters are also estimated by a combined GEE like (5.13) (GEE2); these are Models 3B and 4B.

Table 7.1:

Model	Covariance Structure	GEE1/GEE2	Working Correlation	Description
1A	$\text{cov}(y_{ij},\ y_{kl}) = 0,\ i \neq k\ \&\ j \neq l,$ $\text{var}(y_{ij}) = \phi_j \mu_{ij},\ \forall i, j.$		WC1	Indept.
1B	$\text{var}(y_{ij}) = \phi_j \mu_{ij},\ \forall i, j.$	GEE2	WC2	General
2	$\text{var}(y_{ij}) = \phi_{ij} \mu_{ij},\ \forall i, j.$ $\log(\phi_{ij}) = \epsilon_{0j} + \epsilon_{1j}\text{Base}$	GEE2	WC2	Log-linked Dispersion
3A	$\text{var}(y_{ij}) = \mu_{ij} + \phi_j \mu_{ij}^2$	GEE1	WC2	Quadratic
3B	As Model 3A	GEE2	WC2	Quadratic
4A	$\text{var}(y_{ij}) = \phi_{ij} g(\mu_{ij})$ $\phi_{ij} = 1 + \phi_j \mu_{ij}\ \&\ g(\mu_{ij}) = \mu_{ij}$	GEE1	WC2	Reparam. Quadratic
4B	As Model 4A	GEE2	WC2	Reparam. Quadratic

All models except 1A (Independence) have $\rho_{ijk} = \alpha,\ j \neq k$, (Equal Correlation).

The selection of the final profile model specification is done in stages and the first of the two multistage estimating strategies discussed in §6.2 will be used, because the number of data units is not large. The final forms for the models considered are listed in Table 7.2 and refer to DFTSA models.

The estimates of the parameters and standard errors for Models 3A and 4A are virtually the same and so Model 4A is not shown. This is expected as they are effectively the same model (3A and 4A differ only in the choice of a variance or dispersion response vector respectively) and estimation has been performed by separate GEE in both cases. Models 3A and 4A show no significant evidence of any of the parameters being occasion-specific. However the high parameter estimate at Visit 1 in Table 7.3 indicates some evidence of a higher Age effect at the first time point. Table 7.3 shows the occasion-specific parameter estimates for Age for Model 3A with all other mean parameters fixed over time. The final profile model forms of Models 1A, 1B, 2 and 4B indicate that each of the mean parameters of Trt, Base.Trt and Age significantly differ over time. The main sources of these differences over time for these parameters occur between the first time point and the latter 3 time points, which are not significantly different. This indicates the treatment is having a major effect at the first time point with lower impact afterwards, but older patients respond less favourably to the treatment initially. Thus for Models 1A, 1B, 2 and 4B, the profile component matrix in (6.3) for Age has the form,

$$M_{age} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}.$$

Similarly, the profile component matrices for Trt and Base.Trt have the same form. Using the marginal distribution for the Age estimates $\hat{\boldsymbol{\beta}}_{age}$, and specifying the model $\boldsymbol{\beta}_{age} = \boldsymbol{M}_{age}\boldsymbol{\omega}$ leads to a generalized least squares estimate of $\boldsymbol{\omega}$. A crude goodness of fit

$$(\hat{\boldsymbol{\beta}}_{age} - \boldsymbol{M}_{age}\boldsymbol{\omega})'\boldsymbol{V}_{age}^{-1}(\hat{\boldsymbol{\beta}}_{age} - \boldsymbol{M}_{age}\boldsymbol{\omega})$$

Table 7.2:

Param.	Model					
	1A	1B	2	3A	3B	4B
$\hat{\beta}_0^*$	1.658	1.628	1.557	1.679	1.740	1.703
	(0.084)	(0.083)	(0.085)	(0.072)	(0.069)	(0.071)
$\hat{\beta}_1$	0.931	0.922	1.016	0.886	*	0.876
	(0.084)	(0.083)	(0.078)	(0.101)	*	(0.153)
$\hat{\beta}_{21}$	-2.120	-2.168	-2.122			-1.411
	(0.471)	(0.431)	(0.459)			(0.574)
$\hat{\beta}_2$	-1.178	-1.257	-1.138	-0.982	-0.790	-0.690
	(0.471)	(0.431)	(0.459)	(0.427)	(0.602)	(0.503)
$\hat{\beta}_{31}$	0.907	0.918	0.907			0.544
	(0.177)	(0.178)	(0.163)			(0.257)
$\hat{\beta}_3$	0.469	0.492	0.448	0.357	0.274	0.215
	(0.170)	(0.178)	(0.155)	(0.208)	(0.270)	(0.225)
$\hat{\beta}_{41}$	1.618	1.634	1.741			1.217
	(0.392)	(0.393)	(0.408)			(0.296)
$\hat{\beta}_4$	0.606	0.594	0.561	0.515	*	0.232
	(0.283)	(0.307)	(0.322)	(0.282)	*	(0.267)
$\hat{\alpha}$		0.409	0.425	0.351	0.393	0.374
		(0.061)	(0.060)	(0.101)	(0.073)	(0.072)
$\hat{\phi}_1$	2.944	3.643		0.525	0.455	0.352
	(0.717)	(0.704)		(0.169)	(0.171)	(0.103)
$\hat{\phi}_2$	4.390	4.133		0.530	0.396	0.466
	(1.448)	(1.599)		(0.309)	(0.236)	(0.240)
$\hat{\phi}_3$	8.274	8.358		0.791	0.599	0.806
	(4.002)	(4.390)		(0.495)	(0.247)	(.436)
$\hat{\phi}_4$	1.972	1.960		0.096	0.117	0.111
	(0.303)	(0.389)		(0.050)	(0.052)	(0.052)
$\hat{\epsilon}_0^*$			1.538			
			(0.234)			
$\hat{\epsilon}_1$			-0.218			
			(0.198)			
$\hat{\epsilon}_{13}$			1.195			
			(0.508)			
$\hat{\epsilon}_2$			-0.726			
			(0.290)			
$\log(G)$	35.0	40.0	53.2	41.5	66.8	55.0
AIC	-483.4	-462.6	-450.1	-463.8	-459.3	-459.2
SC	-495.9	-476.1	-463.6	-474.2	-474.9	-472.7
CAIC	-501.9	-482.6	-470.1	-479.2	-482.4	-479.2

* Model 3B estimates for Trt and Age are displayed in Table 7.4.

Table 7.3:

Param.	Visit1	Visit2	Visit3	Visit4
Age	1.075	0.270	0.255	0.476
	(0.354)	(0.338)	(0.430)	(0.327)

Table 7.4:

Trt and Age parameter estimates for Model 3B.						
β_{11}	β_{12}	β_{13}	β_{14}	β_{41}	β_{42}	$\beta_{4(34)}$
0.916	0.685	1.193	0.809	0.792	-0.419	0.155
(0.226)	(0.173)	(0.254)	(0.149)	(0.236)	(0.245)	(0.231)

for this model can be performed, which has an asymptotic χ^2 on 2 d.f. The parameter estimates for $\boldsymbol{\omega}$ (in Model 2) are $(1.752\ 0.410)'$ with estimated standard errors $(0.401\ 0.289)'$, and the goodness of fit statistic is 1.669, indicating a good fit. Similar goodness of fit tests for Trt and Base.Trt produced the values 0.517 and 0.683 respectively, also indicating good fits.

In Model 3B no clear trends or patterns appear. In this model the Base parameter, in contrast with the other models including Model 4B, varies over time ($\chi^2 = 14.6354$, 3 d.f.), and pairwise tests on the four time points of Base show strong differences for different times. For Age, the results are only a little more informative, there being little difference between time points 3 and 4, but differences between other time point combinations. Table 7.4 shows the parameter estimates for Base and Age (combined estimate for time points 3 and 4). These peculiarities between Model 3B and the others may be due to the influence of

$$\partial\boldsymbol{\sigma}_i/\partial\boldsymbol{\beta} = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta} + 2\boldsymbol{\phi}_i^*. * \boldsymbol{\mu}_i. * \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta} \tag{7.2}$$

appearing in $\boldsymbol{D}_i = \partial\boldsymbol{\eta}_i/\partial\boldsymbol{\lambda}$, where here $\boldsymbol{\sigma}_i$ replaces $\boldsymbol{\phi}_i$ for Model 3B and $\boldsymbol{\phi}_i^* =$

$(\phi_1,\ \phi_2,\ \phi_3,\ \phi_4)'$. The binary operator, '.*', multiplies the elements of the left vector operand with the corresponding rows of the right matrix operand. The term $\partial\boldsymbol{\phi}_i/\partial\boldsymbol{\beta}$ is zero in the other models except in Model 4B, where it is

$$\boldsymbol{\phi}_i{}^{*}.*\,\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}, \tag{7.3}$$

and which has lesser impact than $\partial\boldsymbol{\sigma}_i/\partial\boldsymbol{\beta}$ in Model 3B. Thus it is observed that the mean model inferences have changed between GEE1 and GEE2, and for alternative specifications (under GEE2).

There is no indication of a Visit 4 effect in any of the models considered, even though it is significant if included in the mean model with all parameters fixed over time. However in Model 2, the intercept parameters ϵ^{*}_{0j} for the dispersion model significantly differ over time. This difference is due to strong differences between ϵ^{*}_{04} and the other time points, and so is equivalent to a fixed intercept over time and a Visit 4 effect. This observation agrees with the low parameter estimate $\hat{\phi}_4$, as estimated from the other models. As well, there are strong differences between the third time point and the other time points for Base in the dispersion model. This is not surprising because of the large parameter estimate for ϕ_3 in the other models.

Using the $\log(G)$ criterion Model 3B would be considered the best, but it has the largest number of parameters compared to the other models, and meaningful patterns are not obvious for this model. The next best would be Model 4B (which is really equivalent to 3B) followed by Model 2. Models 3A and 4A show the loss of efficiency by using two separate GEE. The advantage of Model 2 is that due to the choice of 'working' covariance matrix and dispersion mean model, the mean parameters are consistent even if the dispersion and correlation models are incorrectly specified.

These conclusions contrast the information provided by AIC and SC (inferences based on SC and CAIC are generally identical), which attempt to account for differing parameter sizes. The ranking of models using AIC and SC differ in the middle range but both select the same 2 models for the first and second positions, these models being Models 2 (first) and 4B (second). Jones' (1993) [61] modified AIC also selects

Model 2 as the best model. Intuitively, modelling the dispersion with a Base variable is not unreasonable. Individuals with large base counts may have large changes in seizure counts as the treatment is applied, tapering off over time, compared to those with smaller base counts with small changes. These changes may appear as a variable dispersion.

As with $\log(G)$, AIC and SC show the inefficiency of Model 1A. Similarly, AIC and Jones' criteria imply (but not as dramatically as $\log(G)$) estimation of Models 3A and 4A by GEE1 results in loss of information compared to Models 3B and 4B and GEE2. Criterion SC is more neutral on this issue, and CAIC selects Model 3A over 3B. The four information criteria are unanimous in showing that the alternative variance specification to Models 3A and 3B (i.e. Model 4B) is either as good as, or usually better. Clearly the variance specification used has a significant effect in the resulting goodness of fit of the model (especially if used in conjunction with GEE2).

Table 7.5 provides a comparison of models where the mean parameters are now all fixed over time and a binary Visit 4 term added. This allows a comparison with the models considered in Thall and Vail (1990) [142]. All the models in Table 7.5 except 2A and 2B are the same as those described for the occasion-specific case in terms of the variance and correlation. In Model 2A the intercept and Base are fixed over time and there is a Visit 4 effect. Model 2B has no Visit 4 term.

The estimates for Model 4A are identical to 3A except the estimated standard errors are slightly lower for the ϕ's and α in 4A. As in Table 7.2, we simplify dispersion Model 2 to an intercept term fixed over time, plus a significant Visit 4 effect. However, there are strong differences between time points 1 and 3 with the other time points for the dispersion base parameter. Model 2B shows a significant base term in the dispersion model, but this is not so in Model 2A.

Models 2, 3B (and 3A) and 4B (and 4A) appear to be the best under the $\log(G)$ model criterion, while not surprisingly Model 1A is by far the worst. Even though Model 2 has the highest value for the goodness of fit test $\log(G)$, its larger number of parameters makes comparison to the other models difficult by this test criterion.

Table 7.5:

Param.	Model							
	1A	1B	2	2A	2B	3A	3B	4B
$\hat{\beta}_0$	-2.717	-2.181	-1.980	-1.588	-1.679	-1.431	-0.559	-1.112
	(0.907)	(0.929)	(1.001)	(0.900)	(0.913)	(0.910)	(0.735)	(0.836)
$\hat{\beta}_1$	0.932	0.920	1.008	0.918	0.913	0.903	0.868	0.907
	(0.086)	(0.084)	(0.077)	(0.100)	(0.120)	(0.097)	(0.182)	(0.148)
$\hat{\beta}_2$	-1.438	-1.356	-1.152	-1.123	-1.052	-0.989	-0.759	-0.801
	(0.415)	(0.422)	(0.370)	(0.430)	(0.428)	(0.419)	(0.599)	(0.507)
$\hat{\beta}_3$	0.595	0.547	0.454	0.439	0.427	0.365	0.232	0.267
	(0.170)	(0.176)	(0.148)	(0.200)	(0.202)	(0.203)	(0.303)	(0.228)
$\hat{\beta}_4$	0.902	0.753	0.615	0.577	0.601	0.541	0.311	0.444
	(0.266)	(0.274)	(0.293)	(0.265)	(0.262)	(0.270)	(0.192)	(0.239)
$\hat{\beta}_5$	-0.168	-0.162	-0.128	-0.153	-0.150	-0.158	-0.147	-0.154
	(0.065)	(0.063)	(0.062)	(0.074)	(0.078)	(0.074)	(0.074)	(0.072)
$\hat{\alpha}$		0.390	0.401	0.367	0.349	0.347	0.390	0.363
		(0.059)	(0.060)	(0.060)	(0.056)	(0.101)	(0.081)	(0.065)
$\hat{\phi}_1$	3.252	4.324				0.391	0.438	0.417
	(0.718)	(0.774)				(0.127)	(0.112)	(0.106)
$\hat{\phi}_2$	4.267	3.733				0.383	0.346	0.368
	(1.372)	(1.292)				(0.230)	(0.181)	(0.173)
$\hat{\phi}_3$	7.436	6.990				0.615	0.615	0.614
	(3.419)	(3.608)				(0.383)	(0.359)	(.316)
$\hat{\phi}_4$	2.189	1.983				0.101	0.112	0.108
	(0.360)	(0.397)				(0.051)	(0.060)	(0.055)
$\hat{\epsilon}_0^*$			1.480					
			(0.216)					
$\hat{\epsilon}_{11}$			0.461					
			(0.369)					
$\hat{\epsilon}_{12}$			-0.255					
			(0.184)					
$\hat{\epsilon}_{13}$			1.138					
			(0.474)					
$\hat{\epsilon}_{14}$			-0.258					
			(0.353)					
$\hat{\epsilon}_0$				0.723	0.321			
				(0.441)	(0.465)			
$\hat{\epsilon}_1$				0.433	0.570			
				(0.280)	(0.273)			
$\hat{\epsilon}_2$			-0.662	-0.773				
			(0.264)	(0.292)				
$\log(G)$	27.8	33.8	51.8	41.8	38.5	49.5	49.2	50.0
AIC	-487.3	-469.4	-457.0	-466.8	-472.1	-463.2	-463.2	-462.7
SC	-497.7	-480.8	-470.5	-477.2	-481.5	-474.6	-474.6	-474.2
CAIC	-502.7	-486.3	-477.0	-482.2	-486.0	-480.1	-480.1	-479.7

The selection criteria AIC and SC agree for the order of ranking for the models in Table 7.5. Further they support the results of $\log(G)$ in terms of the best models, again confirming Model 2 as the best model and Model 4B as the next best. Jones' (1993) [61] modified AIC also indicates Model 2 is the best model. This agrees with the results seen in Table 7.2 for the occasion-specific case. We can also use the selection criteria AIC and SC to provide meaningful comparisons of models between Tables 7.2 and 7.5. Amongst all the models considered in Tables 7.2 and 7.5, Model 2 (Table 7.2) is still by far the best followed by Model 2 (Table 7.5) and then by Model 4B (Table 7.2).

Overall, the best models imply that occasion-specific models may be important and introducing a Base variable in the dispersion better represents dispersion variation. Effectively, as indicated by Models 2A and 2B (but less clearly in 2), the larger the baseline, the greater the impact on the dispersion which does not seem unreasonable. Though not considered here, instead of explicitly modelling the dispersion in terms of the baseline as in Model 2, a random effects model with a baseline random effect could be used. The baseline random variable could be categorical, indicating high, medium or low baseline values.

Under AIC with Jones' (1993) [61] selection proposal, Model 2 (Table 7.2) is the best choice, however Models 2 (Table 7.5), 3B and 4B (Table 7.2) must be considered competitors for second best model. As Model 3B contains the largest number of parameters amongst the three it is excluded, the other two models containing an equal number of parameters. The higher AIC value of Model 2 probably puts it ahead of Model 4B.

In Table 7.6 a selection of the models considered in Table 7.5 is given, but the equal correlation parameter is estimated by a moment estimator (see §6.6). The case of an unstructured covariance matrix (moment estimated) is given in Table 7.7. The best models in Tables 7.6 and 7.7 are the same as Table 7.5, and estimates of the parameters and standard errors are in general quite similar.

The baseline parameter estimate $\hat{\epsilon}_1$ in Models 2A and 2B for Tables 7.6 and 7.7 is significant, and differs little between the two models, providing further support of a

Table 7.6:

Equal Correlation and Moment Correlation Estimator

Param.	Model				
	1B	2	2A	2B	3B
$\hat{\beta}_0$	-2.800 (0.918)	-1.672 (0.925)	-1.218 (0.902)	-1.438 (0.906)	-1.657 (0.808)
$\hat{\beta}_1$	0.920 (0.083)	1.032 (0.070)	.900 (0.107)	0.901 (0.124)	0.888 (0.170)
$\hat{\beta}_2$	-1.547 (0.417)	-0.987 (0.356)	-1.024 (0.427)	-1.003 (0.419)	-0.911 (0.484)
$\hat{\beta}_3$	0.635 (0.171)	0.384 (0.142)	0.392 (0.207)	0.402 (0.203)	0.366 (0.214)
$\hat{\beta}_4$	0.939 (0.271)	0.503 (0.267)	0.476 (0.265)	0.536 (0.260)	0.606 (0.209)
$\hat{\beta}_5$	-0.170 (0.066)	-0.118 (0.066)	-0.149 (0.082)	-0.147 (0.084)	-0.149 (0.079)
$\hat{\phi}_1$	3.183 (0.696)				0.321 (0.103)
$\hat{\phi}_2$	4.353 (1.484)				0.461 (0.223)
$\hat{\phi}_3$	7.442 (3.505)				0.715 (0.348)
$\hat{\phi}_4$	2.215 (0.367)				0.141 (0.046)
$\hat{\epsilon}_0^*$		1.475 (0.212)			
$\hat{\epsilon}_{11}$		0.431 (0.410)			
$\hat{\epsilon}_{12}$		-0.379 (0.232)			
$\hat{\epsilon}_{13}$		1.416 (0.421)			
$\hat{\epsilon}_{14}$		0.083 (0.320)			
$\hat{\epsilon}_0$			0.320 (0.479)	0.082 (0.456)	
$\hat{\epsilon}_1$			0.644 (0.253)	0.699 (0.244)	
$\hat{\epsilon}_2$		-0.697 (0.232)	-0.736 (0.257)		
$\hat{\alpha}$	0.382	0.387	0.355	0.347	0.366
log(G)	27.5	47.0	35.9	32.4	45.0
AIC	-467.5	-455.7	-466.2	-471.4	-461.9
SC	-477.9	-468.1	-475.5	-479.7	-472.2
CAIC	-482.9	-474.1	-480.0	-483.7	-477.1

Table 7.7:

Unstructured Correlation and Moment Correlation Estimator

Param.	Model				
	1B	2	2A	2B	3B
$\hat{\beta}_0$	-3.144 (0.940)	-1.779 (0.948)	-1.706 (0.915)	-1.872 (0.918)	-1.953 (0.808)
$\hat{\beta}_1$	0.920 (0.084)	1.033 (0.069)	0.912 (0.102)	0.919 (0.117)	0.881 (0.163)
$\hat{\beta}_2$	-1.678 (0.419)	-1.086 (0.351)	-1.192 (0.434)	-1.142 (0.418)	-1.017 (0.473)
$\hat{\beta}_3$	0.689 (0.169)	0.416 (0.137)	0.461 (0.207)	0.457 (0.202)	0.411 (0.210)
$\hat{\beta}_4$	1.043 (0.277)	0.533 (0.271)	0.618 (0.270)	0.657 (0.265)	0.700 (0.212)
$\hat{\beta}_5$	-0.163 (0.063)	-0.103 (0.060)	-0.145 (0.080)	-0.143 (0.084)	-0.143 (0.079)
$\hat{\phi}_1$	3.112 (0.694)				0.314 (0.102)
$\hat{\phi}_2$	4.589 (1.641)				0.488 (0.244)
$\hat{\phi}_3$	7.534 (3.486)				0.725 (0.348)
$\hat{\phi}_4$	2.295 (0.391)				0.139 (0.046)
$\hat{\epsilon}_0^*$		1.500 (0.217)			
$\hat{\epsilon}_{11}$		0.409 (0.401)			
$\hat{\epsilon}_{12}$		-0.416 (0.233)			
$\hat{\epsilon}_{13}$		1.387 (0.427)			
$\hat{\epsilon}_{14}$		0.049 (0.315)			
$\hat{\epsilon}_0$			0.461 (0.483)	0.170 (0.084)	
$\hat{\epsilon}_1$			0.674 (0.259)	0.648 (0.253)	
$\hat{\epsilon}_2$		-0.721 (0.255)	-0.762 (0.243)		
$\log(G)$	27.1	47.0	35.8	32.4	44.7
AIC	-463.9	-451.7	-463.3	-468.9	-459.4
SC	-474.3	-464.1	-472.6	-477.2	-469.8
CAIC	-479.3	-470.1	-477.1	-481.2	-474.8

baseline effect in the variance. The baseline estimates in Models 2A and 2B are higher in Tables 7.6 and 7.7 than Table 7.5. Moving to the higher Model 2A from 2B alters the magnitude little for this estimate under moment correlation estimation. This is not the case in Table 7.5 where the parameter estimate is more model sensitive (reduction in magnitude when moving to the upper model) accounting for the non-significance of baseline in Model 2A.

The results in Tables 7.6 and 7.7 further support incorporating baseline into a dispersion model. Unfortunately the parameter estimates of the dispersion baseline covariate of Model 2 displayed in Tables 7.2, 7.5, 7.6 and 7.7 are not in themselves very informative. The individual estimates have high standard errors except at the third time point. This indicates that the third time point contributes greatly to the overall significance of baseline in the dispersion.

Referring to Tables 7.5, 7.6 and 7.7, the standard errors of the mean parameters (except for the intercept and a marginally higher Age) are the lowest for Model 2. This indicates Model 2 has an overall positive effect reducing the standard error of the mean parameters, at least when the mean parameters are fixed over time. These observations do not generally hold in Table 7.2.

The model sensitivity in Table 7.2 demonstrates the difficulty in analysis when incorporating dispersion/variance estimation within the GEE framework. The choice of the model specification can alter the inferences (e.g. Models 3B and 4B) as in the extended quasi-likelihood case. The choice of GEE1 (modified in the context of this example) or GEE2 also affects model sensitivity complicating the issue further. In Models 1A, 1B and 2, GEE1 and GEE2 are the same because the working correlation is WC2 (WC1 for 1A) and $\partial \boldsymbol{\rho}_i / \partial \boldsymbol{\beta} = \mathbf{0}$ (page 85) and $\partial \boldsymbol{\phi}_i / \partial \boldsymbol{\beta} = \mathbf{0}$ (page 100) in $\boldsymbol{D}_i$. Thus the estimate of $\boldsymbol{\beta}$ will be consistent for incorrectly specified correlation and dispersion models. Though GEE1 and GEE2 are not the same under Models 3B and 4B, Model 4B is closer to achieving the pseudo-orthogonal properties of GEE1 than Model 3B. This follows from the contribution of (7.2) and (7.3) to $\boldsymbol{D}_i$ for Models 3B and 4B respectively.

The model sensitivity problem encountered in this example casts some doubt on the reliability or security of using GEE2 here. However choosing model specifications that generate estimating equations which are in some sense closer to the pseudo-orthogonal GEE1, may be an acceptable compromise utilizing all available information on the mean parameters.

The impact of choice of working correlation upon parameter estimation especially in terms of dispersion models needs further study.

Computationally, estimation of the parameters in Table 7.5 presents little difficulty over estimation using a moment correlation estimator in Table 7.6. Moment estimation does allow however a reduction in the complexity of the estimating equations.

7.2.2 Fertility example

The second example involves a study of prolactin levels in women (Archer (1977) [5]) and was examined in Paik (1992) [104]. The prolactin levels were measured at four times, of 15 minute intervals, after an injection of thyrotropin releasing hormone (TRH). A total of 30 subjects were examined forming three groups depending upon the status of their fertility. The subjects groups were classified as normal, infertile with an endometrial biopsy and infertile with lateral deficiency, with group sizes 6, 12 and 12 respectively. Two baseline levels were recorded at -10 and 0 minutes before injection of TRH. The explanatory variables are the indicators for groups 2 (G2) and 3 (G3), time and the average of the two baseline levels. We assume a log link for both the mean and dispersion unless otherwise specified. If the data followed a Gamma distribution, then a constant coefficient of variation (CV) implies a constant dispersion parameter. Table 4 in Paik (1992) [104] shows a slight increase in time for the CV for groups 1 and 2, and a decrease for group 3. The data also displays some heteroscedasticity.

The mean model considered is

$$\log(\mu_i) = \beta_0 + \beta_1 \mathrm{G2} + \beta_2 \mathrm{G3} + \beta_3 \mathrm{Time} + \beta_4 \mathrm{Base},$$

and the dispersion models (assuming correlated within subject responses except Model

1A) are displayed in Table 7.8.

The subscript k in Model 2 refers to the group number. Model 3 uses the empirical variances for the response vector in (6.7) in a similar manner to Prentice and Zhao (1991) [110]. Model 3A is a reparameterization of Model 3, and the dispersion response vector is used for estimation.

The working correlations for the mean responses either follow an Equal correlation structure such as that in Model 1B for the seizure data, or an AR-1 structure. All models considered are in conjunction with WC2 except for 1A which uses WC1.

The ϕ parameter in Model 1B can also be estimated using an empirical variance response vector in a combined GEE or from two separate GEE (like Model 3A of the seizure example). These models are denoted 1C and 1D respectively in Table 7.8.

Paik (1992) [104] estimates the mean and dispersion parameters by (5.7) and (6.7) separately and uses an AR-1 structure for both the mean and dispersion. The AR-1 parameters are then estimated by moment estimators. Model 1B is equivalent to variance Model 5 in Table 5 of Paik (1992) [104] and Model 4 and variance Model 1 are the same.

The estimates for Models 1-5 are displayed in Tables 7.9 and 7.10 for Equal and AR-1 correlation structures respectively. Model 1A appears in Table 7.9 only. Convergence problems for Models 1B and 1D for the AR-1 case result in the exclusion of these models from Table 7.10. Model 1D yields identical estimates as Model 1B in Table 7.9 except the standard error of α is more than twice that of Model 1B. Interestingly Models 1A and 1B (Table 7.9) produce identical estimates for the mean and dispersion parameters and their respective standard errors. Only the goodness of fit statistics $\log(G)$, AIC, SC and CAIC signify the inefficiency of assuming complete independence (Model 1A). The goodness of fit tests SC and CAIC as expected produced identical inferences for the fertility data.

A test of equality of the dispersion parameters for the 3 groups in Model 2 results in the acceptance of this hypothesis for both Equal and AR-1 correlation structures.

Table 7.8:

Model	Covariance Structure	GEE1/GEE2	Working Correlation	Description
1A	$\text{cov}(y_{ij},\ y_{kl}) = 0,\ i \neq k\ \&\ j \neq l,$ $\text{var}(y_{ij}) = \phi\mu_{ij}^2,\ \forall i, j.$		WC1	Indept.
1B	$\text{var}(y_{ij}) = \phi_{ij} g(\mu_{ij}),\ \forall i, j.$ $\phi_{ij} = \phi\ \&\ g(\mu_{ij}) = \mu_{ij}^2$	GEE2	WC2	General
1C	$\text{var}(y_{ij}) = \phi\mu_{ij}^2,\ \forall i, j.$	GEE2	WC2	Reparam. General
1D	As Model 1C	GEE1	WC2	Reparam. General
2	$\text{var}(y_{ij}) = \phi_{ij} g(\mu_{ij}),\ \forall i, j.$ $\phi_{ij} = \phi_k,\ k = 1, 2, 3\ \&$ $g(\mu_{ij}) = \mu_{ij}^2$	GEE2	WC2	Group
3	$\text{var}(y_{ij}) = \phi_1\mu_{ij}^2 + \phi_2\mu_{ij}^3$	GEE2	WC2	Cubic
3A	$\text{var}(y_{ij}) = \phi_{ij} g(\mu_{ij})$ $\phi_{ij} = \phi_1\mu_{ij} + \phi_2\mu_{ij}^2\ \&$ $g(\mu_{ij}) = \mu_{ij}$	GEE2	WC2	Reparam. Cubic
4	$\text{var}(y_{ij}) = \phi_{ij}\mu_{ij},\ \forall i, j.$ $\log(\phi_{ij}) = \epsilon_0 + \epsilon_1\text{G3} + \epsilon_2\text{Time}+$ $\epsilon_3\text{G3.Time} + \epsilon_4\text{G1.Time}$	GEE2	WC2	Log-linked Dispersion
5	$\text{var}(y_{ij}) = \phi_{ij}\mu_{ij},\ \forall i, j.$ $\log(\phi_{ij}) = \epsilon_0 + \epsilon_1\text{G3} + \epsilon_2\text{Time}+$ $\epsilon_3\text{G3.Time} + \epsilon_4\text{G1.Time} + \epsilon_5\text{Base}$	GEE2	WC2	Log-linked Dispersion

Table 7.9:

Equal Correlation Structure								
Param.	Model							
	1A	1B	1C	2	3	3A	4	5
$\hat{\beta}_0$	4.468 (.187)	4.468 (.187)	4.522 (.358)	4.471 (.188)	4.567 (.325)	4.554 (.176)	4.478 (.185)	4.493 (.186)
$\hat{\beta}_1$	-.104 (.191)	-.104 (.191)	-.208 (.421)	-.103 (.192)	-.190 (.563)	-.183 (.246)	-.089 (.190)	-.095 (.195)
$\hat{\beta}_2$	.425 (.209)	.425 (.209)	.256 (.508)	.427 (.210)	.170 (.839)	.193 (.274)	.372 (.217)	.424 (.212)
$\hat{\beta}_3$	-.265 (.017)	-.265 (.017)	-.276 (.028)	-.265 (.017)	-.269 (.016)	-.268 (.016)	-.263 (.017)	-.268 (.017)
$\hat{\beta}_4$	.005 (.008)	.005 (.008)	.011 (.008)	.005 (.008)	.008 (.017)	.008 (.007)	.003 (.008)	.002 (.008)
$\hat{\alpha}_e$		.830 (.050)	.836 (.044)	.829 (.044)	.842 (.115)	.840 (.042)	.837 (.041)	.833 (.065)
$\hat{\phi}$	.143 (.027)	.143 (.027)	.146 (.029)					
$\hat{\phi}_{G1}$				.171 (.137)				
$\hat{\phi}_{G2}$				.136 (.045)				
$\hat{\phi}_{G3}$				.135 (.031)				
$\hat{\phi}_1$					.087 (.268)	.090 (.123)		
$\hat{\phi}_2$					.001 (.004)	.001 (.002)		
$\hat{\epsilon}_0$							-1.875 (.436)	-1.967 (.932)
$\hat{\epsilon}_1$							.638 (.750)	.191 (.787)
$\hat{\epsilon}_2$							-.025 (.233)	-.057 (.295)
$\hat{\epsilon}_3$							-.250 (.299)	-.153 (.285)
$\hat{\epsilon}_4$							.073 (.434)	.081 (.778)
$\hat{\epsilon}_5$								.015 (.029)
log(G)	38.3	45.3	44.0	56.3	58.4	59.7	54.1	62.8
AIC	-429.7	-369.8	-369.9	-371.4	-371.5	-371.2	-371.8	-372.9
SC	-433.9	-374.7	-374.8	-377.7	-377.1	-376.8	-379.5	-381.3
CAIC	-436.9	-378.2	-378.3	-382.2	-381.1	-380.8	-385.0	-387.3

Table 7.10:

AR-1 Correlation Structure						
Param.	Model					
	1C	2	3	3A	4	5
$\hat{\beta_0}$	4.416 (.156)	4.441 (.180)	4.405 (.264)	4.396 (.314)	4.448 (.179)	4.439 (.179)
$\hat{\beta_1}$	.035 (.140)	-.076 (.185)	.037 (.224)	.009 (.346)	-.094 (.184)	-.087 (.182)
$\hat{\beta_2}$	.384 (.070)	.438 (.201)	.416 (.620)	.427 (.458)	.445 (.201)	.416 (.204)
$\hat{\beta_3}$	-.252 (.048)	-.263 (.017)	-.259 (.025)	-.260 (.020)	-.263 (.018)	-.262 (.018)
$\hat{\beta_4}$	.003 (.004)	.005 (.008)	.004 (.010)	.005 (.007)	.004 (.008)	.005 (.008)
α_a	.873 (.041)	.880 (.037)	.870 (.049)	.869 (.042)	.879 (.046)	.877 (.054)
$\hat{\phi}$	.131 (.023)					
$\hat{\phi_{G1}}$		.136 (.011)				
$\hat{\phi_{G2}}$		.174 (.024)				
$\hat{\phi_{G3}}$		.115 (.033)				
$\hat{\phi_1}$			.149 (.218)	.154 (.051)		
$\hat{\phi_2}$			-.0003 (.0034)	-.0004 (.0007)		
$\hat{\epsilon_0}$					-2.020 (.619)	-1.968 (.730)
$\hat{\epsilon_1}$					-.181 (.891)	-.001 (1.421)
$\hat{\epsilon_2}$					.115 (.200)	.130 (.211)
$\hat{\epsilon_3}$					-.106 (.329)	-.153 (.672)
$\hat{\epsilon_4}$					-.107 (.092)	-.116 (.094)
$\hat{\epsilon_5}$						-.009 (.039)
log(G)	51.2	63.7	64.9	63.6	62.7	71.7
AIC	-363.5	-364.4	-364.6	-364.4	-366.3	-367.4
SC	-368.4	-370.7	-370.2	-370.0	-374.0	-375.8
CAIC	-371.9	-375.2	-374.2	-374.0	-379.5	-381.8

Similarly the parameter ϕ_2 in Models 3 and 3A is not significant. In none of the models was the baseline variable shown to be significant, in contrast to Paik (1992) [104], due to the higher standard errors reported here. None of the parameters in the dispersion Models 4 and 5 appear to be important, though G3.Time was significant in Model 1 of Paik. The standard errors recorded for the dispersion parameters in Model 4 for both AR-1 and Equal Correlation are generally much greater than those of Paik's Model 1, though the mean parameters are similar except for the baseline.

We note that even though Models 4 and 5 are rated strongly under $\log(G)$ for both AR-1 and Equal Correlations, Model 5 being the best, AIC and SC clearly demonstrate they are the worst (after Model 1A). This supports the non-significant nature of the parameters in the dispersion models and highlights the misleading nature of $\log(G)$ especially for small data sets. With such a small sample size it is obvious that little information can be obtained from the variance. Therefore it is not surprising AIC and SC lean towards selecting Model 1B (and consequently 1C) as the best and simplest model.

One further result is the higher efficiency of the AR-1 correlation to Equal Correlation for this data set. This is strongly supported by the four goodness of fit statistics when comparing dispersion/variance models in Table 7.9 with the same models in Table 7.10. This is reflected in the generally lower standard errors (except for Model 3A) of the mean parameters under AR-1.

The models in Tables 7.9 and 7.10 were repeated with the 'working covariance' submatrix $\boldsymbol{V}_{dr}$ set to the $\mathbf{0}$ matrix, that is the vectors $\boldsymbol{d}$ and $\boldsymbol{r}$ have zero covariance matrix. In the case of Model 1B for example, this is equivalent to estimating $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ by three separate GEE, that is a GEE1 type estimation. Tables 7.11 (Equal Correlation) and 7.12 (AR-1) list the estimators for the re-estimated models. Models 5A and 5B in Tables 7.11 and 7.10 are submodels of Model 5 (and Model 4). Insignificant terms are gradually removed in Model 5 leading to Model 5B.

For Equal Correlation, $\log(G)$ is higher in Table 7.11 ($\boldsymbol{V}_{dr} = \mathbf{0}$) than in Table 7.9.

Table 7.11:

Equal Correlation Structure ($V_{dr} = 0$)									
Param.	Model								
	1B	1C	2	3	3A	4	5	5A	5B
$\hat{\beta}_0$	4.468 (.187)	4.497 (.165)	4.480 (.189)	4.490 (.165)	4.477 (.170)	4.507 (.184)	4.579 (.172)	4.570 (.188)	4.470 (.188)
$\hat{\beta}_1$	-.104 (.191)	-.158 (.167)	-.099 (.192)	-.158 (.168)	-.135 (.174)	-.101 (.194)	-.147 (.193)	-.122 (.204)	-.111 (.193)
$\hat{\beta}_2$	.425 (.209)	.369 (.186)	.433 (.211)	.355 (.185)	.379 (.190)	.425 (.218)	.585 (.189)	.548 (.194)	.442 (.207)
$\hat{\beta}_3$	-.265 (.017)	-.266 (.016)	-.266 (.017)	-.264 (.017)	-.265 (.017)	-.266 (.018)	-.278 (.017)	-.276 (.016)	-.266 (.017)
$\hat{\beta}_4$	.005 (.008)	.006 (.006)	.004 (.008)	.006 (.006)	.006 (.006)	-.0003 (.0078)	-.008 (.005)	-.005 (.006)	.004 (.008)
$\hat{\alpha_e}$	.830 (.050)	.830 (.052)	.828 (.054)	.831 (.052)	.831 (.050)	.840 (.049)	.824 (.066)	.820 (.066)	.830 (.048)
$\hat{\phi}$	.143 (.027)	.143 (.027)							
$\hat{\phi}_{G1}$			.189 (.079)						
$\hat{\phi}_{G2}$			.139 (.050)						
$\hat{\phi}_{G3}$			.126 (.046)						
$\hat{\phi}_1$				.127 (.045)	.128 (.048)				
$\hat{\phi}_2$				.0003 (.0005)	.0003 (.0005)				
$\hat{\epsilon}_0$						-2.028 (.268)	-2.651 (.322)	-2.260 (.296)	-2.047 (.390)
$\hat{\epsilon}_1$						.532 (.475)	-.400 (.427)	-1.007 (.525)	
$\hat{\epsilon}_2$						.112 (.082)	.119 (.085)		
$\hat{\epsilon}_3$						-.339 (.157)	-.267 (.152)		
$\hat{\epsilon}_4$						.018 (.139)	.068 (.135)		
$\hat{\epsilon}_5$							.059 (.020)	.050 (.019)	.007 (.023)
log($\hat{G}$)	45.3	46.2	59.9	61.9	61.6	60.0	70.5	53.8	51.1
AIC	-369.7	-369.5	-371.6	-370.3	-370.4	-371.6	-373.6	-372.3	-370.7
SC	-374.7	-374.4	-377.9	-375.9	-376.0	-379.3	-382.0	-378.6	-376.3
CAIC	-378.2	-377.9	-382.4	-379.9	-380.0	-384.8	-388.0	-383.1	-380.3

Table 7.12:

AR1 Correlation Structure ($V_{dr} = 0$)									
Param.	Model								
	1B	1C	2	3	3A	4	5	5A	5B
$\hat{\beta}_0$	4.400 (.180)	4.424 (.151)	4.388 (.180)	4.420 (.149)	4.409 (.158)	4.410 (.173)	4.482 (.167)	4.473 (.182)	4.402 (.181)
$\hat{\beta}_1$	-.090 (.183)	-.145 (.148)	-.094 (.183)	-.145 (.147)	-.122 (.158)	-.097 (.180)	-.123 (.185)	-.104 (.195)	-.098 (.185)
$\hat{\beta}_2$	.411 (.202)	.368 (.174)	.403 (.203)	.356 (.171)	.375 (.178)	.382 (.205)	.525 (.184)	.527 (.190)	.429 (.199)
$\hat{\beta}_3$	-.259 (.017)	-.261 (.015)	-.258 (.017)	-.258 (.017)	-.259 (.017)	-.258 (.017)	-.270 (.017)	-.268 (.016)	-.260 (.017)
$\hat{\beta}_4$	.008 (.008)	.009 (.006)	.009 (.008)	.009 (.005)	.009 (.006)	.007 (.007)	-.0004 (.0054)	.0008 (.0064)	.008 (.008)
$\hat{\alpha_a}$	.894 (.051)	.896 (.051)	.898 (.054)	.895 (.050)	.895 (.050)	.903 (.051)	.894 (.063)	.889 (.065)	.894 (.049)
$\hat{\phi}$	.143 (.028)	.143 (.028)							
$\hat{\phi_{G1}}$			.181 (.073)						
$\hat{\phi_{G2}}$			.124 (.047)						
$\hat{\phi_{G3}}$			.143 (.049)						
$\hat{\phi_1}$				.124 (.046)	.125 (.048)				
$\hat{\phi_2}$				.0003 (.0006)	.0003 (.0006)				
$\hat{\epsilon_0}$						-2.138 (.286)	-2.606 (.331)	-2.236 (.297)	-2.065 (.366)
$\hat{\epsilon_1}$						.725 (.446)	-.048 (.443)	-.717 (.532)	
$\hat{\epsilon_2}$						.067 (.084)	.092 (.079)		
$\hat{\epsilon_3}$						-.279 (.135)	-.260 (.138)		
$\hat{\epsilon_4}$						.075 (.146)	.102 (.135)		
$\hat{\epsilon_5}$							.047 (.017)	.040 (.019)	.008 (.021)
log(G)	45.1	46.4	57.3	62.0	61.5	60.3	71.4	53.7	50.8
AIC	-362.2	-362.3	-364.8	-363.4	-363.4	-365.6	-369.6	-365.9	-363.7
SC	-362.2	-362.3	-364.8	-363.4	-363.4	-365.6	-369.6	-372.2	-369.3
CAIC	-370.6	-370.7	-375.6	-373.0	-373.0	-378.8	-384.0	-376.7	-373.3

This reflects the fact that $\log(G)$ is a function of the variance of $\boldsymbol{\lambda}$ and may be sensitive to different working covariance matrices. The AIC, SC and CAIC values have changed little, indicating no preference for either table.

The AR-1 case is interesting because there is a drop in $\log(G)$ values in Table 7.12 relative to Table 7.10. This is in contrast to what was observed for Equal Correlation and indicates some interacting effect of the 'working covariance' matrix for the second order response variables and the correlation model for the measured response variables. The other test statistics, AIC, SC and CAIC indicate little change.

Comparing Tables 7.11 and 7.12 indicates a closing of the gap between Equal Correlation and AR-1 in terms of $\log(G)$. In fact closer examination shows some of the $\log(G)$ values under Equal Correlation have overtaken their AR-1 counterparts. The other goodness of fit criteria, including Jones' modified AIC criterion support the superiority of AR-1. Thus $\log(G)$ may indeed be a less robust method than other methods which tend to be reasonably consistent even though the distribution used is not technically correct (since we are using a Normal approximation).

Estimated standard errors for the mean parameters of Models 3 and 3A are much lower when $\boldsymbol{V}_{dr} = \mathbf{0}$ is used. The same applies to the standard errors of the dispersion parameters in Models 4 and 5. The Baseline parameter becomes quite significant here compared to Tables 7.9 and 7.10. This significance is maintained down to Model 5A but disappears rather dramatically in Model 5B.

The question that arises is, whether there is any real advantage in setting $\boldsymbol{V}_{dr} = \mathbf{0}$. A major problem with this example is the small sample size and the large sample asymptotics are probably not very appropriate. This would at least partially explain the difficulty encountered in fitting the models in Tables 7.9 and 7.10 for a number of cases. Setting $\boldsymbol{V}_{dr} = \mathbf{0}$ has reduced the instability in estimation, especially for AR-1, and the number of cycles required for convergence. The standard errors between all models in Tables 7.11 and 7.12 are similar, in contrast with Tables 7.10 and 7.9 where standard errors can differ significantly. Using $\boldsymbol{V}_{dr} = \mathbf{0}$ may possibly result in standard errors that

are more realistic. Also in a number of cases, $V_{dr} = \mathbf{0}$ is equivalent to estimating all the parameters by three separate GEE leading to consistent mean estimates irrespective of the dispersion/variance and correlation models chosen. As specified earlier, consistency of the GEE estimates is upheld in the face of incorrect specification of the covariance model. The discussion in §5.6 also highlighted concerns for the estimation of the mean parameters if inappropriate working models for the third and fourth moments were used. These considerations have led to setting $V_{dr} = \mathbf{0}$ as a valid modification for WC2 difficulties.

Tables 7.9-7.12 highlight the danger in basing sole goodness of fit criterion on $\log(G)$, and the information criteria provide useful supplementary information on the goodness of fit.

The models in Tables 7.11 and 7.12 were re-estimated with the correlation parameters estimated by moment estimators. In general the mean and dispersion parameter estimates and standard errors were either identical or very close to the corresponding estimates listed in Tables 7.11 and 7.12. The moment estimates of the Equal Correlation parameter were virtually identical with their counterparts in Table 7.11. The AR-1 correlation parameter was higher however, ranging from 0.922 to 0.937. As was seen in Tables 7.11 and 7.12, the $\log(G)$ criterion showed little difference between Equal Correlation and AR-1. These differences were slightly in favour of AR-1 (from 0.2 to 0.8 above). The other goodness of fit statistics demonstrated a more noticeable, but not overly striking leaning towards AR-1 (1 to 2 units above except for a few cases where higher differences were observed).

Table 7.13 represents the models of Tables 7.11 and 7.12 with respect to the mean and dispersion parameters, and the 'working covariance' matrix is the general unstructured case. The unstructured $\log(G)$ values are very similar to those of AR-1 and Equal Correlation (α estimated by the method of moments). The other goodness of fit statistics are, as expected, much more favourable to the unstructured case. The correlation matrix

Table 7.13:

Unstructured Correlation Structure							
Param.	Model						
	1B	2	3	4	5	5A	5B
$\hat{\beta}_0$	4.419 (.177)	4.418 (.177)	4.425 (.147)	4.451 (.170)	4.535 (.161)	4.524 (.174)	4.428 (.177)
$\hat{\beta}_1$	-.117 (.177)	-.117 (.177)	-.154 (.139)	-.136 (.179)	-.176 (.183)	-.139 (.190)	-.127 (.179)
$\hat{\beta}_2$	.398 (.198)	.397 (.199)	.355 (.192)	.394 (.201)	.542 (.179)	.547 (.182)	.424 (.194)
$\hat{\beta}_3$	-.272 (.017)	-.272 (.017)	-.272 (.017)	-.272 (.017)	-.284 (.017)	-.282 (.016)	-.272 (.017)
$\hat{\beta}_4$	.008 (.008)	.008 (.008)	.010 (.007)	.005 (.007)	-.003 (.005)	-.002 (.006)	-.008 (.008)
$\hat{\phi}$	.148 (.029)						
$\hat{\phi}_{G1}$		.172 (.074)					
$\hat{\phi}_{G2}$		.139 (.051)					
$\hat{\phi}_{G3}$		.144 (.051)					
$\hat{\phi}_1$			.132 (.046)				
$\hat{\phi}_2$			.0003 (.0006)				
$\hat{\epsilon}_0$				-2.118 (.275)	-2.625 (.332)	-2.272 (.281)	-2.084 (.362)
$\hat{\epsilon}_1$				.676 (.461)	-.171 (.432)	-.866 (.535)	
$\hat{\epsilon}_2$				.132 (.087)	.148 (.091)		
$\hat{\epsilon}_3$				-.338 (.157)	-.298 (.174)		
$\hat{\epsilon}_4$				-.032 (.147)	.013 (.133)		
$\hat{\epsilon}_5$					.051 (.016)	.049 (.017)	.012 (.021)
log(G)	38.4	50.3	54.9	53.3	64.6	47.7	44.5
AIC	-358.5	-360.4	-359.7	-361.1	-363.4	-362.1	-360.1
SC	-362.7	-366.0	-364.6	-368.1	-372.1	-367.7	-365.0
CAIC	-365.7	-370.0	-368.1	-373.1	-377.6	-371.7	-368.5

estimate for Model 1B in Table 7.13 is

$$\begin{bmatrix} 1 & 0.875 & 0.748 & 0.723 \\ & 1 & 0.911 & 0.858 \\ & & 1 & 0.875 \\ & & & 1 \end{bmatrix},$$

with similar values for the other models. The standard errors in Table 7.13 are comparable to the corresponding models in Tables 7.11 and 7.12.

An adjustment to WC2 can be made to reflect Gamma distributed type data by letting the variance of d_{ij} equal $2\phi_{ij}^2(1+3\phi_{ij})$ (Paik (1992) [104]) rather than $2\phi_{ij}^2$ as has been used here. Alternatively, rather than using fourth moments derived by assuming the response vectors $\boldsymbol{y}_i$ are Normally distributed, I suggest a 'working covariance' matrix developed from a multivariate Gamma distribution, such as that described in Johnson and Kotz (1972) [60] (pp. 216-220), could be used.

As for the seizure case, the choice of model specification can strongly affect the outcome of estimation, particularly for Equal Correlation and 'working correlation' WC2 (Table 7.9). Under Model 1B, GEE1 and GEE2 are the same (with respect to $\boldsymbol{\beta}$) because of the choice of WC2. Recall that the estimates under Model 1B and Model 1D were very close. The main difference with Model 1D (pseudo-orthogonal GEE1) is the much lower standard error of α under Model 1B, possibly reflecting the loss of information in using GEE1 for Model 1D. The standard errors under the pseudo-orthogonal Model 1B are much lower than Model 1C. Model 3A is closer to fulfilling the pseudo-orthogonality properties implied by GEE1 than the equivalent Model 3. This is reflected by the much higher standard errors of Model 3 displayed in Table 7.9. Overall for Equal Correlation, the choice of the specification is important under GEE2. Under AR-1, the standard errors under Models 3 and 3A (see Table 7.10) are equally high, relative to the other models. The AR-1 case differs from the Equal Correlation case in that the most successful model out of 1B, 1C and 1D (in the sense that it was the only one that converged) was Model 1C (GEE2).

When $\boldsymbol{V}_{dr}$ is set to $\mathbf{0}$ (as in Tables 7.11 and 7.12), the standard errors are lower for Models 1C and 3 than Models 1B and 3A respectively. The magnitude of difference (of the standard errors) is much lower though than in Table 7.9. This reversal for Equal Correlation (compared to when $\boldsymbol{V}_{dr} \neq \mathbf{0}$) may occur because the dispersion response vector involves dividing by the estimate of the assumed variance function $g(\mu_{ij})$ (as in Model 1B) compared to the empirical covariance response vector (as used in Model 1C). This could result in a greater degree of imprecision in the overall estimating procedure. Further, to some extent $\boldsymbol{V}_{dr} = \mathbf{0}$ reduces the non-orthogonality of Models 1C and 3 (and 3A under AR-1) with GEE2 compared to Tables 7.9 and 7.10 negating much of the impact of $\partial\boldsymbol{\sigma}_i/\partial\boldsymbol{\beta}$ (or $\partial\boldsymbol{\phi}_i/\partial\boldsymbol{\beta}$) in $\boldsymbol{D}_i$.

This example reinforces the orthogonality considerations discussed in the seizure example (at least for Equal Correlation), but it also highlights the effect of 'working correlations' on the estimation particularly under GEE2. A poor choice of the 'working covariance' for the first and second order GEE's may seriously handicap the estimating procedure, especially for small data sets. The particular choice of the variance/dispersion specification is also an important consideration. At least for this example, extending the GEE estimation to the correlation parameters may not be the best approach compared to moment estimation which provides a robust and stable procedure.

7.2.3 SPRINT example

The third example considered is data from the Secondary Prevention Reinfarction Israeli Nifedipine Trial (SPRINT) data reported in Laor and Cohen (1992) [73]. The aim of the study was to examine the efficacy of Nifedipine in preventing myocardial infarction (MI), and included 2320 patients from 13 heart institutes in Israel. Patients were randomly allocated a placebo or the drug, forming two groups, within 7 to 21 days after their MI. The data included one 24 hour electocardiogram (ECG) recording for each patient taken 3-6 months after MI.

Laor and Cohen (1992) [73] examined the hourly count over the 24 hours of ventric-

ular premature heartbeats (VPB) as a response variable. They considered a set of 31 explanatory variables, which were all binary categorical variables, before arriving at a final set of significant variables with estimates displayed in their Table 1. Patients were only considered if complete ECG recordings and explanatory variables were available (906 patients). Comparisons between the group where full data was available and the group where it was not, implied that it was not unreasonable to consider the final data set of 906 patients as representative of the patient population of interest. Finally, only patients with at least one VPB non-zero were considered, giving a sample size of 686 patients. The VPB counts were combined into four 6 hour totals.

The final set of variables included an intercept term (Int), an indicator variable for the first time block (I1), and the second time block (I2) and a ANT and Age interaction term (ANT.Age). The ANT is the measured evidence of anterior MI and Age is a binary variable that indicates < 55 or not. The explanatory variables, effort onset angina pectoris with functional capacity III or IV (EFANS), and MI according to ECG (MI) were also included in the final model. The relationship of the mean and variance was taken to be $g(\mu_{ij}) = \mu_{ij}^2$ where the variance is (5.6) with a log-link function for the mean.

Two data sets were made available by Laor and Cohen (1992) [73]. One of the data sets was composed of the full 906 patients with VPB counts grouped into four 6 hour totals. The other data set appears to consist of a subset of the 686 (at least one non-zero count) patients and complete hourly counts for the 24 hours were available. The size of this subpopulation was $n_s = 423$. The VPB counts for this subset were grouped into eight 3 hour blocks. Indicator variables were created for each of the first four blocks (I11, I12, I21 and I22). A mean model with indicators representing the first 2 time blocks (I1) and the second 2 (I2) corresponding to the four 6 hour blocks case was also fitted. The 2 models for the mean are

Model 1 ($t = 8$):

$$\log(\mu_i) = \beta_0 + \beta_{11}\mathrm{I11} + \beta_{12}\mathrm{I12} + \beta_{21}\mathrm{I21} + \beta_{22}\mathrm{I22} + \beta_3\mathrm{ANT.Age} + \beta_4\mathrm{EFANS} + \beta_5\mathrm{MI},$$

Model 2 ($t = 4,\ 8$):

$$\log(\mu_i) = \beta_0 + \beta_1 \mathrm{I1} + \beta_2 \mathrm{I2} + \beta_3 \mathrm{ANT.Age} + \beta_4 \mathrm{EFANS} + \beta_5 \mathrm{MI}.$$

Models 1 and 2 are considered for the subset data ($t = 8$ time points), and Model 1 for the full data ($t = 4$). Both models were used with WC2.

A difficulty arose for the full 906 patient data set, and which casts some doubt over the other data set made available. Estimators of some of the mean parameters and standard errors, and particularly the dispersion parameter, for the 686 sub-population patients examined here do not agree with those of Table I of Laor and Cohen (1992) [73]. We have endeavoured to reproduce the results of their paper but have not been successful. Out of the 906 patients in our copy of their data set, there were indeed 686 patients with at least one non-zero count, and the outlier, $(2800,\ 0,\ 0,\ 2)'$, reported in their paper was indeed present.

Table 7.14 reproduces Table II (except the fitted values column) in Laor and Cohen (1992) [73], and represents the distribution of the VPB by subgroups (covariate profiles) with the outlier removed. The distribution of the VPB by subgroups for our copy of this data set is listed in Table 7.15. The percentiles in Table 7.15 were obtained by the quantile function in the Splus statistical package which uses a linear interpolation between quantiles. The counts of subgroups, N, are the same for both tables, but the mean VPB and percentiles of the VPB differ. Some of the 99% percentile values in Table 7.14 are greater than the 100% percentile values in Table 7.15 for the same subgroups. This indicates that at least in these cases, the maximum values from Table 7.14 for these subgroups are higher than the same subgroups or profiles in our data. These two data sets which should be the same somehow differ. Attempts to contact Laor and Cohen to this date have been unsuccessful. Given this lack of communication, we can only make the assumption that our data is correct or free of contamination.

The eigenvalues and eigenvectors of the sample correlation matrices for the entire data (t=4) and its subset (t=8) were examined. For $t = 4$ the eigenvalues were

$$\boldsymbol{e}_1 = (0.1624\ 0.2582\ 0.4052\ 3.1742)',$$

Table 7.14:

Distribution of VPB by subgroups (Laor & Cohen)										
							Percentiles of VPB			
MID	EV	AA	EF	MI	*N*	Mean VPB	50	75	90	99
0	0	0	0	0	26	60.38	2	26	165	765
0	0	0	0	1	922	95.90	3	36	170	1815
0	0	0	1	0	12	68.58	46	112.5	130	277
0	0	0	1	1	136	265.93	9	144	710	3050
0	0	1	0	1	238	20.32	2	9	49	274
0	0	1	1	1	36	16.32	0.5	3	69	212
0	1	0	0	0	13	30.15	0	12	149	165
0	1	0	0	1	461	136.23	4	47	247	2210
0	1	0	1	0	6	57.83	26.5	105	187	187
0	1	0	1	1	68	414.22	9	227.5	1130	5500
0	1	1	0	1	119	31.11	3	15	59	510
0	1	1	1	1	18	8.83	1	4	26	91
1	0	0	0	0	13	21.13	3	12	72	130
1	0	0	0	1	461	120.30	4	38	213	1990
1	0	0	1	0	6	77.0	40.5	87	294	294
1	0	0	1	1	68	409.69	16	243.5	1040	5150
1	0	1	0	1	119	27.28	3	18	61	464
1	0	1	1	1	18	22.22	1.5	12	132	155

Table 7.15:

Distribution of VPB by subgroups											
							Percentiles of VPB				
MID	EV	AA	EF	MI	*N*	Mean VPB	50	75	90	99	100
0	0	0	0	0	26	51.00	1	11.25	127.5	615	765
0	0	0	0	1	922	124.40	3	38	203.9	2155.3	7700
0	0	0	1	0	12	55.67	28.5	105.75	116.1	179.3	187
0	0	0	1	1	136	320.18	9	121.25	842.5	4522.5	5500
0	0	1	0	1	238	28.56	2	12.75	60.2	477.07	960
0	0	1	1	1	36	9.72	1	4	22.5	99.45	104
0	1	0	0	0	13	21.23	3	12	66.4	123.04	130
0	1	0	0	1	461	120.30	4	38	213	1972	4750
0	1	0	1	0	6	77	40.5	85.25	190.5	283.65	294
0	1	0	1	1	68	409.69	16	224.25	1034.4	4513.5	5150
0	1	1	0	1	119	27.28	3	16.5	57	432.68	470
0	1	1	1	1	18	22.22	1.5	9.75	90	151.09	155
1	0	0	0	0	13	48.92	5	26	137.4	364.76	392
1	0	0	0	1	461	79.19	3	38	174	1504	2020
1	0	0	1	0	6	83.67	47	119.75	203.5	269.65	277
1	0	0	1	1	68	305.72	9	216.25	768.5	3193.88	3486
1	0	1	0	1	119	14.62	1	8	37	168.56	274
1	0	1	1	1	18	22.22	0	2.5	57	204.69	212

with eigenvectors

$$\boldsymbol{E}_1 = \begin{bmatrix} -0.0573 & 0.7722 & -0.3920 & 0.4967 \\ 0.7425 & -0.3764 & -0.1907 & 0.5203 \\ -0.6668 & -0.4917 & -0.2168 & 0.5164 \\ -0.0291 & 0.1424 & 0.8735 & 0.4647 \end{bmatrix}.$$

If we assume an Equal Correlation matrix, then using our preliminary estimate $\alpha = 0.7276$ for the correlation parameter, the resulting eigenvalues of the Equal Correlation matrix was

$$\boldsymbol{e}_2 = (0.2724\ 0.2724\ 0.2724\ 3.1827)',$$

which is similar to $\boldsymbol{e}_1$, and the eigenvector corresponding to the largest eigenvalue is 0.51_4, where $\mathbf{1}_4$ is a unit vector of size 4. This is very close to the last eigenvector in $\boldsymbol{E}_1$. The eigenvalues of the sample correlation matrix of the subset data ($t = 8$) were

$$\boldsymbol{e}_3 = (0.2280\ 0.1884\ 0.1032\ 0.0541\ 0.4751\ 0.6359\ 0.8604\ 5.4439)',$$

and the eigenvector of the largest eigenvalue (5.4439) is

$$\boldsymbol{E}_{3(8)} = (0.3805\ 0.3708\ 0.3886\ 0.3974\ 0.3389\ 0.3567\ 0.3223\ 0.2507).'$$

Again assuming an Equal Correlation parameter with our preliminary estimate $\alpha = 0.6283$, the eigenvalues of the Equal Correlation matrix are 0.3747 (multiplicity 7) and 5.3774 (multiplicity 1) which are similar to $\boldsymbol{e}_3$. The eigenvector corresponding to the largest eigenvalue (5.3774) is 0.35361_8, which is very close to $\boldsymbol{E}_{3(8)}$. Such observations provide strong evidence that an Equal Correlation matrix would provide a good approximation for both data sets.

The form of the inverse covariance/correlation matrix for an Autoregressive process for a Normal distribution was also used. The inverse of the covariance matrix is derived in Verbyla (1985) [145] and is shown in equation (3) of that paper. The inverse of the correlation matrix follows directly. The inverse of the sample correlation matrix is then compared to this form. In the case of the subset data ($t = 8$), an AR-1 approximation

was found to be entirely inadequate. With the entire data ($t = 4$), it was not as clear whether the inverse of the sample correlation matrix had a form similar to the inverse of an AR-1 matrix (Normal approximation). Examination of the estimated unstructured correlation matrix (with and without outlier) in Table 1 of Laor and Cohen (1992) [73] implies the AR-1 structure is probably not warranted.

Setting $\boldsymbol{V}_{dr} = \mathbf{0}$ for both data sets resulted in identical parameter estimates as their full WC2 counterparts listed in Table 7.16. Unfortunately Model 2 (AR-1) did not converge (except for $\boldsymbol{V}_{dr} = \mathbf{0}$) and may be attributable to the unsuitability of the AR-1 case. The goodness of fit statistics AIC, SC and CAIC are weighted towards an AR-1 model despite the earlier evidence to the contrary. Since models have been considered for both data sets (one is a subset of the other), comparisons of models by the goodness of fit statistics between the two data sets is not appropriate. The information criteria favour Model 1 with AR-1 for the subset data ($t = 8$) in Table 7.16. Under Equal Correlation, Model 2 ($t = 8$) is approximately as good as Model 1 ($t = 8$) using AIC and SC (given the magnitude of the observed criteria).

Since the number of parameters are identical between models with the same mean and dispersion structure but different correlation structures, the large sample suggests the asymptotics of the parameter estimates are more valid. This implies that $\log(G)$ may provide a reliable statistic to base inferences. Further, Figure 2 in Laor and Cohen (1992) [73] displays the histogram of VPB's for the four time blocks, and which shows the data as being highly skewed. This makes the Normality approximation used in AIC, SC and CAIC highly suspect and conclusions based on these statistics very dubious indeed.

There is little difference between the Equal or AR-1 correlation structures by the $\log(G)$ statistic. The standard errors for the mean parameters are either marginally lower (except for β_0 and β_5 ($t = 8$) and β_1 ($t = 4$) which are marginally higher) or significantly lower (β_3 and $t = 8$) for the Equal Correlation structure. The AR-1 structure reduces the size of the standard error of the dispersion parameter however.

Table 7.16:

Param.	Model					
	$t = 8$				$t = 4$	
	Model 1 (Eq.)	Model 1 (AR1)	Model 2 (EQ)	Model 2 (AR1 & $V_{dr} = 0$)	Model 2 (EQ)	Model 2 (AR1)
$\hat{\beta}_0$	3.368 (.462)	3.459 (.450)	3.368 (.461)	3.476 (.448)	3.714 (.475)	3.895 (.493)
$\hat{\beta}_1$			0.079 (.114)	-0.029 (.159)	-0.215 (.190)	-0.267 (.188)
$\hat{\beta}_2$			0.171 (.110)	0.050 (.114)	-0.008 (.072)	-0.068 (.078)
$\hat{\beta}_{11}$	0.086 (.119)	0.040 (.137)				
$\hat{\beta}_{12}$	0.072 (.123)	0.084 (.165)				
$\hat{\beta}_{21}$	0.175 (.116)	0.084 (.121)				
$\hat{\beta}_{22}$	0.167 (.114)	0.046 (.121)				
$\hat{\beta}_3$	-1.264 (.292)	-1.006 (.453)	-1.264 (.292)	-1.010 (.456)	-1.416 (.309)	-1.314 (.363)
$\hat{\beta}_4$	0.868 (.281)	0.986 (.297)	0.867 (.281)	0.991 (.298)	0.883 (.290)	0.859 (.292)
$\hat{\beta}_5$	1.071 (.456)	0.891 (.453)	1.071 (.456)	0.900 (.443)	1.082 (.477)	0.882 (.497)
$\hat{\alpha}$	0.375 (.095)	0.818 (.043)	0.375 (.095)	0.834 (.037)	0.456 (.087)	0.621 (.049)
$\hat{\phi}$	11.927 (2.660)	13.895 (1.825)	11.927 (2.660)	13.733 (1.729)	15.829 (2.661)	18.651 (2.230)
log(G)	36.2	36.0	25.5	25.8	25.0	24.9
AIC	-20260	-20065	-20258	-20141	-17308	-17205
SC	-20280	-20085	-20274	-20157	-17326	-17223
CAIC	-20285	-20090	-20278	-20161	-17330	-17227

Table 7.17 displays Models 1 ($t = 8$) and 2 ($t = 4$) examined in Table 7.16, but now the correlation parameters are estimated by moment estimators. Also listed are the models evaluated with an unstructured correlation matrix. Model 3 represents a separate dispersion model. The estimated unstructured correlation matrix for Model 2 ($t = 8$) is

$$\begin{bmatrix} 1 & 0.789 & 0.784 & 0.737 & 0.586 & 0.647 & 0.478 & 0.157 \\ & 1 & 0.758 & 0.656 & 0.418 & 0.688 & 0.529 & 0.142 \\ & & 1 & 0.873 & 0.746 & 0.636 & 0.504 & 0.120 \\ & & & 1 & 0.769 & 0.699 & 0.478 & 0.127 \\ & & & & 1 & 0.693 & 0.385 & 0.089 \\ & & & & & 1 & 0.750 & 0.144 \\ & & & & & & 1 & 0.200 \\ & & & & & & & 1 \end{bmatrix},$$

and ($t = 4$)

$$\begin{bmatrix} 1 & 0.407 & 0.440 & 0.328 \\ & 1 & 0.710 & 0.465 \\ & & 1 & 0.506 \\ & & & 1 \end{bmatrix}.$$

Notice that an Equal Correlation for $t = 8$ may be adequate if the last column is ignored. The estimate of the last column (the correlation of the 8^{th} time block with the other time blocks) is indeed unusual. The results given in Table 7.17 indicate a different conclusion to that of Table 7.16 in that the AIC, SC and CAIC are strongly in favour of Equal Correlation compared to AR-1. This is further supported by $\log(G)$. As expected, the AIC, SC and CAIC definitely favour the unstructured case, but more importantly the $\log(G)$ favours Equal Correlation over the Unstructured case when $t = 8$, and is very close when $t = 4$. Interrelated with the $\log(G)$ results is the observation that the standard errors for the mean parameters are lower for Equal Correlation than for AR-1 in Table 7.17. Further, the standard errors for Equal Correlation and the Unstructured case for Model 1 ($t = 8$) and Model 2 ($t = 8$) are generally similar (except Equal Correlation

Table 7.17:

Param.	Model								
	$t = 8$			$t = 4$					
	Model 1 (Eq.)	Model 1 (AR1)	Model 1 (Unstr.)	Model 2 (Eq.)	Model 2 (AR1)	Model 2 (Unstr.)	Model 3 (Eq.)	Model 3 (AR1)	Model 3 (Unstr.)
$\hat{\beta}_0$	3.368	3.447	3.391	3.714	4.020	3.864	3.584	3.315	3.780
	(.462)	(.451)	(.467)	(.475)	(.504)	(.483)	(.569)	(.803)	(.542)
$\hat{\beta}_1$				-0.215	-0.317	-0.239	-0.170	-0.288	-0.187
				(.190)	(.192)	(.179)	(.225)	(.342)	(.204)
$\hat{\beta}_2$				-0.008	-0.093	-0.047	0.013	-0.058	-0.028
				(.072)	(.086)	(.074)	(.078)	(.078)	(.075)
$\hat{\beta}_{11}$	0.086	-0.009	0.066						
	(.119)	(.156)	(.118)						
$\hat{\beta}_{12}$	0.072	-0.045	0.083						
	(.123)	(.189)	(.113)						
$\hat{\beta}_{21}$	0.175	0.052	0.099						
	(.116)	(.133)	(.111)						
$\hat{\beta}_{22}$	0.167	0.028	0.082						
	(.114)	(.131)	(.106)						
$\hat{\beta}_3$	-1.264	-0.902	-0.789	-1.416	-1.233	-1.292	-1.685	-1.966	-1.481
	(.292)	(.526)	(.423)	(.309)	(.413)	(.378)	(.313)	(.648)	(.280)
$\hat{\beta}_4$	0.868	1.041	1.060	0.883	0.843	0.878	0.808	0.896	0.837
	(.281)	(.309)	(.311)	(.292)	(.296)	(.298)	(.295)	(.344)	(.283)
$\hat{\beta}_5$	1.071	0.818	0.891	1.082	0.743	0.908	1.244	1.534	0.979
	(.456)	(.448)	(.466)	(.477)	(.508)	(.486)	(.566)	(.788)	(.541)
$\hat{\phi}$	11.927	14.772	14.570	15.829	17.544	16.884			
	(2.660)	(1.984)	(2.644)	(2.661)	(2.636)	(2.576)			
$\hat{\phi}_1$							34.886	70.111	29.224
							(24.233)	(88.608)	(12.008)
$\hat{\phi}_2$							10.324	12.603	12.049
							(1.294)	(2.862)	(1.861)
$\hat{\phi}_3$							16.720	18.222	17.825
							(5.083)	(7.007)	(6.075)
$\hat{\phi}_4$							11.797	13.975	12.020
							(2.122)	(4.262)	(1.908)
$\hat{\alpha}$	0.504	0.929		0.456	0.881		0.479	0.824	
log(G)	29.9	27.5	28.3	18.6	17.7	18.4	10.9	5.7	11.4
AIC	-20306	-22261	-19145	-17313	-18254	-16966	-17262	-18256	-16928
SC	-20232	-22277	-19161	-17327	-18268	-16979	-17284	-18277	-16950
CAIC	-20327	-22282	-19166	-17331	-18272	-16983	-17289	-18282	-16955

has a significantly lower β_3 standard error). However, the standard errors are lower under the Unstructured case for Model 3. Overall, Equal Correlation performs as well as or better than (except for Model 3) the naive Unstructured or fully parameterized case. All the standard errors under Model 3 and AR-1 are quite poor compared to Equal Correlation and the Unstructured Case.

With respect to Model 3 in Table 7.17, the test for differences among the dispersion parameters leads to the retention of the null hypothesis of no significant differences among the dispersion parameters. The poor fit by this model is clearly indicated by the low $\log(G)$ statistic and the overall higher standard errors for all three correlation structures. Unfortunately, relative to Model 2, this is not verified by the other goodness of fit criteria, again questioning their reliability for this example. A 'separate' dispersion model was also tried for $t = 8$ under Equal, AR-1 and Unstructured correlation matrices but there was little evidence of separate dispersions.

Comparing Models 1 ($t = 8$) and 2 ($t = 4$) in Table 7.16 to the corresponding models in Table 7.17, it is observed that for Equal Correlation, the mean and dispersion parameter estimates and their standard errors are the same (to three decimal points). With AR-1, the parameter estimates are comparable across tables, but generally the full GEE models in Table 7.16 have lower standard errors. The GEE correlation estimates are lower than the corresponding moment estimates (except for Model 2 ($t = 4$) where they are the same to three decimal points). On the one hand, full GEE estimation of all parameters increases the complexity of the estimating procedure, introduces possible stability problems, and has had no effect on the mean and dispersion parameter estimates (and their standard errors) under Equal Correlation. On the other hand, the possible bias and high inefficiency of moment estimates is well known, and full GEE can reduce the standard errors as occurred for AR-1.

The dispersion parameter estimates, $\hat{\phi}$, recorded here are vastly different to those of Laor and Cohen (1992) [73] (with outlier: $\hat{\phi} = .6519$, and without outlier: $\hat{\phi} = .7746$) and may be attributable to a recording error. There are no significant differences between

the four 3 hour block indicators of Model 1 using the χ^2 tests described earlier. What difference there is, is mainly attributable to differences between the first 6 hour block and the second. The indicators are themselves not significant. The lack of significance for the indicator variables I1 and I2 for Model 2 ($t = 4$ and $t = 8$) also occurs though Laor and Cohen (1992) [73] report a significant indicator for the second 6 hour block. The other mean parameter estimates for Models 1 and 2 are reasonably similar to those of Laor and Cohen (1992) [73] with nearly all having marginally lower standard errors than their estimates (with outlier case).

In this example, it was observed that incorporating the correlation estimation into the GEE framework can have a positive effect, reducing the overall standard errors, at least for large data sets where asymptotic properties may be achieved. The trade-off is that the estimation procedure may become computationally complex and unstable and the gains marginal. This example also demonstrates that the choice of 'working correlation' can improve stability and reduce standard errors. Simpler correlation structures such as the Equal Correlation in this example can perform as effectively and more efficiently than the naive Unstructured case. Since our earlier analysis tended towards Equal Correlation, and Table 7.17 supports this, then there is evidence to support that the AR-1 Correlation structure is inferior to Equal Correlation. In terms of the goodness of fit, given that the number of parameters for competing models do not differ greatly, and the sample size is large, it is my belief that $\log(G)$ can be a useful tool. The generally more robust information criteria (AIC, SC and CAIC) based on a multivariate Normal approximation did not appear to perform as well due to the highly skewed nature of the data.

7.3 Discussion

Fitting occasion-specific mean models can highlight key features or trends for the data in question. The occasion-specific parameters themselves may follow some simple relationship over time allowing simplifications to be made, that is a profile specification of

the form (6.3). For example, there is evidence of higher treatment effectiveness, especially for younger patients, at the first time point for the seizure data. Such modelling of the occasion-specific parameters is especially important for small data sets such as the seizure data because as the number of time points and/or covariates becomes large, a potentially very large number of parameters need to be estimated, diluting the power for each variable. The complexity of the model increases and analysis becomes confused.

As was seen for Model 2 in the seizure example, occasion-specific parameters can easily be extended to the dispersion model. A dispersion model provides an intuitive and powerful way to explore and understand the underlying variance mechanism, accounting for heterogeneity and overdispersion. For example, with the seizure data, a high base count individual could have a strong response (compared to a small response for a low base count individual) when the treatment is initially applied, but with diminishing effect over time. This variation could be modelled through a dispersion model. The GEE's for the dispersion model are also relatively simple to implement.

An important benefit of a good dispersion model is in reducing the standard errors of the estimates of all or most of the mean model parameters (e.g. Model 2 in Tables 7.5, 7.6 and 7.7). Hopefully this reduction is due to a good representation of the true variance process rather than some artifact of the dispersion model. The dispersion model may also indicate problems with mean model itself.

As demonstrated for the occasion-specific mean model (seizure data), the choice of the particular dispersion/variance specification can lead to different inferences on the mean under GEE2. This may indicate that the occasion-specific mean models are especially sensitive to the incorrect specification of the dispersion/variance model, that is the mean of the dispersion/variance response vector. Different inferences were also observed for the same model under GEE1 and GEE2 (e.g. Models 3A and 3B in Table 7.2). Stability problems and higher standard errors were also encountered depending on the particular dispersion/variance specification and GEE strategy (GEE1 or GEE2) used.

When the within-unit 'working correlation' was the Equal Correlation, lower standard errors and fewer iterative steps were often encountered in the first two examples for a particular choice of the dispersion/variance specification under GEE2. The specification that usually achieved the above was the one that was in some sense closest to fulfilling the pseudo-orthogonal properties of GEE1. This can generally be judged by how close $\partial\boldsymbol{\sigma}_i/\partial\boldsymbol{\beta}$ or $\partial\boldsymbol{\phi}_i/\partial\boldsymbol{\beta}$ is to $\mathbf{0}$. The pseudo-orthogonal property is important because it ensures consistency of the mean parameters irrespective of whether the mean of the second order response vectors has been correctly defined. However, as the fertility data clearly demonstrate, the effect of the 'working covariance' matrix on the estimation procedure can be quite important as well. It appears that inferences on the mean regression parameters using GEE's can be quite sensitive to the choice of the 'working correlation' matrix. Thus for the sake of robustness, choosing dispersion or variance models, along with a sensible 'working' correlation matrix of the second order responses that potentially increases consistency of the mean parameters is desirable. The potential loss of efficiency of the mean regression parameters for inappropriate 'working covariance' matrices for the second order response variables was discussed in §5.6.

The stability problems (Models 1B and 1D in Table 7.10, fertility example) for the AR-1 'working correlation', contrary to the Equal Correlation case, demonstrated a specification that did not have the desired pseudo-orthogonal property (Model 1C instead of 1B) was the only one to converge. Model 1D (i.e. Model 1C with GEE1) also failed to converge. Thus another factor that needs to be considered is the choice of the correlation design for the first order response vector. As for Equal Correlation, the choice of the 'working covariance' of the second order response vectors was important as well (Models 3 and 3A in Tables 7.10 and 7.12).

In some cases, such as the fertility data, estimating the correlation parameters by moment estimators can lead to an improvement in stability and speed of estimation. The correlation parameters are usually nuisance parameters, and their estimation by GEE's increases computational complexity, often for little gain. However, GEE estimation of

the nuisance parameters can improve the estimation of mean parameters (lower standard errors) as indicated by the AR-1 case in the SPRINT data. Thus full GEE estimation of all parameters should at least be considered. This raises the issue of whether to employ GEE1 or GEE2. Overall (but not always) GEE1 is the safest method, leading to consistent estimates, acknowledging however the possible loss of efficiency to the less robust GEE2. As discussed earlier, if GEE2 is used, choosing a dispersion/variance specification (if one is available) that makes GEE2 closer to GEE1 may be helpful.

The above discussion so far, highlights the need of further study into a number issues. Proposed future work will be to perform a series of simulation studies. These will examine the impact of incorrect specification of the dispersion/variance model on the estimation of the mean parameters under GEE1 and GEE2. The effect of the choice of alternative dispersion/variance specification on the mean parameter estimates and the stability of GEE1 and GEE2 will also be addressed. Similarly, the effect of the choice of 'working correlation' on the estimating procedure is to be considered.

The goodness of fit statistic $\log(G)$ appears to produce spurious conclusions for the seizure and fertility examples. This is because this statistic tends to favour models with a larger number of parameters. Thus caution is recommended in using $\log(G)$ as the sole goodness of fit criterion. Since $\log(G)$ is a function of the asymptotic variance of the estimators, the large sample size of the SPRINT data (implying greater confidence of the asymptotics of the GEE estimates being achieved) indicates $\log(G)$ is a reasonable goodness of fit statistic in this case. The model selection criteria AIC, SC and CAIC using a multivariate Normal approximation are useful and valuable alternative methods to the $\log(G)$ statistic. However, for a highly skewed data set, such as the SPRINT data, AIC, SC and CAIC does not appear to work very well. This is not surprising because of the use of the multivariate Normal approximation. An alternative multivariate approximation to the Normal approximation is to use a multivariate Gamma distribution (Johnson and Kotz (1972) [60] (pp. 216-220)). The above proposed simulation study could be extended to examine the different model selection criteria under

different distributional assumptions.

Chapter 8

Patterned Correlation Matrices and GEE

8.1 Introduction

In many situations a particular covariance/correlation structure or pattern is suggested by the experimental design or physical properties of repeated measurements. In the Normal case, incorporation of the patterned correlation matrix into the estimating procedure and tests of hypotheses on particular patterns is fairly straightforward. However, in the non-Normal case, including the correlation matrix into the estimation procedure is no longer a simple affair.

The GEE approach provides a method to model data with particular patterned correlation matrices. It provides a natural and straightforward way to analyse such data analogous to Normal theory, and hypotheses on various correlation structures may be explored. An example of nested correlation structures is considered in relation to rat teeth data.

Table 8.1:

Diet	Cariogenic Control	Trt. A	Trt. B
1	No		
2	Yes		
3	Yes	0.25%	
4	Yes	0.50%	
5	Yes	1.00%	
6	Yes		0.25%
7	Yes		0.50%
8	Yes		1.00%

8.2 A Description of the Rat Teeth data

The rat teeth data are found in Table 43.1 of Andrews and Herzberg (1985) [4] (pp. 245-248), and supplied by the General Food Corporation. A total of 120 rats were randomly assigned to 8 diets. The experimental design is displayed in Table 8.1. For example Diet 1 is noncariogenic control (NC), Diet 2 is cariogenic control (CC) and Diet 3 applies 0.25% of Trt A to CC. The aim of the experiment was to see if Treatments A and B would reduce the cariogenic effects of Diet 2.

Three rats died before the completion of the study, two in Diet 3 and one in Diet 7. At the completion of the trial, the rats were killed and their teeth removed and stained. A total of 28 occlusal surfaces in each rat were examined for caries and severity of decay scores were recorded. The ordinal scores were 0, 1 and 2 for no decay, decay into enamel and decay into dentin respectively.

Unfortunately, the authors or their source do not have any further information on this data set. The site location for each measurement is unknown. However the data itself and the dental structure of a rat contains information that can be used.

Information on the dental structure of rats may be obtained from a number of sources such as Farris and Griffith (1949) [39], Rowett (1957) [123] or Hebel and Stromberg (1986) [56]. A summary of the dental structure of a rat is provided here. A rat has no premolars or canines, only 12 molars and 4 incisors. The incisors occur as pairs at the front of the jaw, one pair on the lower jaw and the other on the upper jaw. Molars occur in 4 rows, 3 large molars per row. The actual dental formula of the rat is thus,

$$I_1^1;\ C_0^0;\ PM_0^0;\ M_3^3 = 16,$$

where I, C, PM and M refer to incisors, canines, premolars and molars respectively. Rows will be labelled here as TR, TL, BR and BL (top-right, top-left, bottom-right and bottom-left respectively). The lower pair of incisors protrudes further into the mouth cavity than the upper pair. Incisors grow continuously throughout the life of the rat, and wear of the incisors normally keeps pace with growth. The implication of the conical shape and rapid wear of the incisors is that they are not conducive to caries (decay) formation. However the form and function of rat molars is similar to those of human molars and thus lend themselves readily to experimental studies in caries. The size of the molars decreases from the first molar (M_1) to the third molar (M_3). Each molar is molariform, that is each molar has several cusps, the tips of which are free from the enamel. The number of cusps per molar is

$$M_1\frac{5}{4}\ M_2\frac{4}{4}\ M_3\frac{3}{3}$$

where the numerator and denominator refer to the upper and lower jaws respectively. The masticatory surface of the molars contain transversely oriented (mainly) enamel folds and tubercles. Three main transverse (four in lower M_2) folds can usually be counted in the adult. A large interdental gap (the diastema) exists between the front incisors and the molars.

In dentistry, an occlusion is the meeting of the teeth of the upper and lower jaws when closed.

It is presumed one measurement is taken on the surface of each incisor and two per molar at a point where the incisor/molar contacts the corresponding incisor/molar on the

other jaw. This would account for the 28 measurements per rat. The two measurements per molar as compared to one per incisor would reflect the greater importance of the molar site in examining caries formation. Obviously some other design structure may have been used but the one assumed seems to be a reasonable one.

Presumably the measurements recorded reflect some systematic experimental design or pattern. Close examination of the occlusal scores highlights structure present in the data. The first 2 measurements per rat appear closely related and differ very strongly from adjacent measurements. Again, at the fifteenth and sixteenth measurement, a close relationship between the pair is evident. This would be logical if the 2 pairs correspond to the incisors which are spatially apart from the molars. The two pairs generally have lower occlusal scores than the trailing adjacent measurements (but admittedly the preceding adjacent measurements to the second pair are generally low as well). The lower scores for these pairs would be expected if they did indeed correspond to the incisor pairs because of the greater resistance to caries formation for incisors.

The remaining measurements probably correspond to measurements on the within-molar sites. Two recording strategies for these measurements seem equally likely. The first assumes measurements 3 to 14 correspond to the top jaw and 17 to 24 to the lower jaw or vice versa. After the first pair (the presumed incisors) the next pair are assumed to be measurements on the same molar but different within-molar sites, and so on until the next pair of presumed incisors. The same molar pattern would follow after the second pair of incisors. The second recording strategy assumes measurements 3 to 14 correspond to the left (or right) side of the mouth and 17 to 28 to the other side. The first 6 measurements for both sides belong to the top (or bottom) jaw and the next 6 per side belong to the bottom (or top) jaw.

It seems reasonable to presume that decay would be asymmetric between the top and bottom jaw but symmetric between the left and right side of the mouth. Such an assumption would provide information on which recording strategy was more likely. Preliminary analysis indicates the second strategy supports the symmetry assumptions

best.

Analysis will be focussed on the 24 molar measurements due to their commonality. The 4 incisor measurements per rat are excluded from the analysis. This is because the incisors differ from the molars by location, physical structure and purpose.

8.3 The Mean Model

To reduce the complexity in this data set, the number of categories are reduced from 3 to 2. The first two categories are amalgamated into a no/minor cavity category with score 0. The last category now has score 1, and is referred to as the severe cavity case.

Let μ_{ij} indicate the mean of the response variable Y_{ij}, $i = 117$ and $j = 24$. It is reasonable to expect differences for severity of decay between the top and bottom jaw. Possibly decay would be greater on the bottom than on the top. The position of the molars within a row may also be important. Decay may increase towards the rear of the jaw. Let factor TB take the value 0 if a measurement belongs to the top jaw, otherwise it is 1. Let factor MOL take the values 0, 1 or 2 corresponding to the first, second or third molar. A very general linear predictor model which includes treatment, TB and MOL main effects and two-way interactions will be considered. The linear predictor model is

$$\eta_{ij} = \beta_0 + [(\mathrm{CC} + \mathrm{A} + \mathrm{B}) * (\mathrm{TB} + \mathrm{MOL}) + (\mathrm{TB} + \mathrm{MOL})\text{^}2](i, j), \qquad (8.1)$$

using the corner-point parameterization such as used by the GLIM program. The '$*$' is a binary operator, indicating the main effects of the linear predictor model are the terms in both arguments of this operator, and that there are two-way interactions involving the terms of one argument with the terms of the second argument. The operator '()^2' is a unary operator indicating two-way interactions between all the terms within its argument. The link function between the mean and the linear predictor is taken to be logistic. The variance of the response Y_{ij} is presumed to be $\phi\mu_{ij}(1 - \mu_{ij})$.

Symmetry implies differences between the left and right side of the mouth should

not on average be large. Similarly within-molar differences would not be expected to be large. However the expanded linear predictor model,

$$\begin{aligned}\eta_{ij} &= \beta_0 + [(\text{CC} + \text{A} + \text{B}) * (\text{TB} + \text{LR} + \text{MOL} + \text{WM}) + \\ &\quad (\text{TB} + \text{LR} + \text{MOL} + \text{WM})\text{^}2](i,j), \end{aligned} \tag{8.2}$$

will be considered as well, where LR and WM are indicator variables of left-right and within-molars positions.

The simple dichotomous case was considered, but analysis of the polytomous case follows by considering the vectors of the indicators $\boldsymbol{y}^*_{ij} = (I(y_{ij} = 0),\ I(y_{ij} = 1),\ I(y_{ij} = 2))'$ (see Liang, Zeger and Qaqish (1992) [82]). Then if $E(y^*_{ijk}) = \pi_{ijk}$, $k = 1,\ 2,\ 3$, use $\text{var}(y^*_{ijk}) = \phi^*\pi_{ijk}(1 - \pi_{ijk})$ and $\text{cov}(y_{ijl},\ y_{ijm}) = -\pi_{ijl}\pi_{ijm}$.

8.4 Nested Correlation Structures

A number of nested correlation structures are proposed. As the size of the within-rat molar measurements is large the number of covariance/correlation parameters that need to be estimated can be quite large (276 for the fully parameterized case). Patterned correlation matrices may significantly reduce the number of parameters.

In the following description, the rat data response matrix has been modified. The columns of the 117×24 response matrix $\boldsymbol{Y}$ are split into the blocks TL, TR, BL and BR each of size 117×6. This is a reordering of the columns of the original data in Table 43.1 of Andrews and Herzberg (1985) [4] (the original presumed molar pattern was TL, BL, TR and BR).

It is reasonable to assume the TL, TR, BL and BR blocks have the same pattern. Denote the common 6×6 correlation matrix for these blocks as $\boldsymbol{\Sigma}_1$. Let TL_i represent the i^{th} measurement for block TL, and similarly for TR_i, BL_i and TR_i. Further suppose

$$\text{corr}(\text{TL}_i,\ \text{TR}_j) = \text{corr}(\text{TL}_j,\ \text{TR}_i),\ \forall\ i,\ j.$$

The 6×6 correlation matrix, $\boldsymbol{\Sigma}_2$ say, between blocks TL and TR is consequently symmetric. Similarly, let the correlation matrices between blocks TL and BL, and between

TL and BR be the symmetric matrices $\boldsymbol{\Sigma}_3$ and $\boldsymbol{\Sigma}_4$. The symmetry argument can be extended: for example the correlation between blocks TR and BR is also $\boldsymbol{\Sigma}_3$. The working correlation matrix becomes

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_2 & \boldsymbol{\Sigma}_3 & \boldsymbol{\Sigma}_4 \\ & \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_4 & \boldsymbol{\Sigma}_3 \\ & & \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_2 \\ & & & \boldsymbol{\Sigma}_1 \end{bmatrix}. \tag{8.3}$$

We call this pattern Model 1. This particular choice of correlation pattern reduces the number of unique correlation parameters to 78. Model 1 is simplified if $\boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_4$, that is the correlation pattern between TL (BL) and TR (BR) is the same as between TL (BL) and BR (TR). This simplification results in 57 unique correlation parameters, and is labelled Model 2. Note that $\boldsymbol{R}$ is the patterned correlation matrix discussed in Olkin (1974) [103] if $\boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \boldsymbol{\Sigma}_4$.

Another simplification of Model 1 that adds more structure is

$$\boldsymbol{\Sigma}_j = \begin{bmatrix} \rho_{j1} & \rho_{j2} & \rho_{j3} & \rho_{j4} & \rho_{j3} & \rho_{j4} \\ & \rho_{j1} & \rho_{j4} & \rho_{j3} & \rho_{j4} & \rho_{j3} \\ & & \rho_{j1} & \rho_{j2} & \rho_{j3} & \rho_{j4} \\ & & & \rho_{j1} & \rho_{j4} & \rho_{j3} \\ & & & & \rho_{j1} & \rho_{j2} \\ & & & & & \rho_{j1} \end{bmatrix}, \; j = 2,\ 3,\ 4. \tag{8.4}$$

This incorporates a number of features:

(i) Correlations of the same sites between rows are the same, i.e. ρ_{j1}.

(ii) The correlation between the first within-molar site and the second within-molar site of the equivalent molar on another row is ρ_{j2}.

(iii) The correlation between the first (second) within-molar site and the first (second) within-molar site is ρ_{j3}, where the molars of the different rows occupy different sites within the row.

(iv) The correlation between the first (second) within-molar site and the second (first) within-molar site is ρ_{j4}, where the molars of the different rows occupy different sites within the row.

Call this structure Model 3, and let Model 4 be the equivalent simplification of Model 2. Such a model implies that the first within-molar site occupies an approximately equivalent spatial location for all molars. The same applies to the second within-molar site. The number of unique correlation parameters reduces to 27 and 23 for Models 3 and 4 respectively.

Naturally $\boldsymbol{\Sigma}_1$ could also be simplified to pattern (8.4), where $\rho_{11} = 1$. Define Models 5 and 6 to be Models 3 and 4 respectively with $\boldsymbol{\Sigma}_1$ having pattern (8.4). This results in a very economical 15 and 11 unique correlation parameters respectively.

The above proposed correlation patterns may be written as a sum of kronecker product matrices. For example, Model 2 can be written as

$$\boldsymbol{R} = \boldsymbol{I}_4 \otimes \boldsymbol{\Sigma}_1 + \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \otimes \boldsymbol{\Sigma}_2 + \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \otimes \boldsymbol{\Sigma}_3.$$

A highly structured correlation matrix that will also be considered is

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho_2 & \rho_1 & \rho_4 \\ \rho_2 & 1 & \rho_4 & \rho_1 \\ \rho_1 & \rho_4 & 1 & \rho_2 \\ \rho_4 & \rho_1 & \rho_2 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & \alpha & \alpha \\ \alpha & 1 & \alpha \\ \alpha & \alpha & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & \rho_3 \\ \rho_3 & 1 \end{bmatrix}. \tag{8.5}$$

The first matrix of (8.5) models the correlation between top and bottom and left and right. The second matrix is the between molar correlation contribution for a given row, with equal correlations applied. The within molar correlation is given by the third matrix. This kronecker form is labelled Model 7. Thus the original 276 unique correlations in the fully parameterized case can be reduced to only 5 parameters. We

Table 8.2:

Model	Correlation Pattern	Comments	$\Sigma_2 = \Sigma_4$
1	Equation (8.3)		No
2	Equation (8.3)		Yes
3	Equation (8.3)	Σ_2, Σ_3 and Σ_4 have structure (8.4)	No
4	Equation (8.3)	Σ_2 and Σ_3 have structure (8.4)	Yes
5	Equation (8.3)	Σ_1, Σ_2, Σ_3 and Σ_4 have structure (8.4)	No
6	Equation (8.3)	Σ_1, Σ_2 and Σ_3 have structure (8.4)	Yes
7	Equation (8.5)		No
8	Equation (8.5)	$\rho_2 = \rho_4$	Yes
9	Equation (8.6)		No

can simplify (8.5) by setting $\rho_4 = \rho_2$, that is correlations between TL and BR, and between BL and TR are the same as between TL and TR, and between BL and BR. This is referred to as Model 8.

Alternatively the kronecker form (8.5) can be extended to

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & \rho_2 \\ \rho_2 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & \alpha & \alpha \\ \alpha & 1 & \alpha \\ \alpha & \alpha & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & \rho_3 \\ \rho_3 & 1 \end{bmatrix}, \tag{8.6}$$

where the first two matrices separately model the correlations between the top and bottom, and between left and right. Refer to this pattern as Model 9.

Figure 8.1 shows the nesting of the correlation patterns that are proposed in this section. Table 8.2 presents a summary of Models 1-9.

8.5 Estimation of the Correlation Parameters

Estimation of the unstructured or fully parameterized correlation matrix is by the moment estimator (6.13). Denote $\boldsymbol{R}_u$ as the matrix of moment estimates of the correlations.

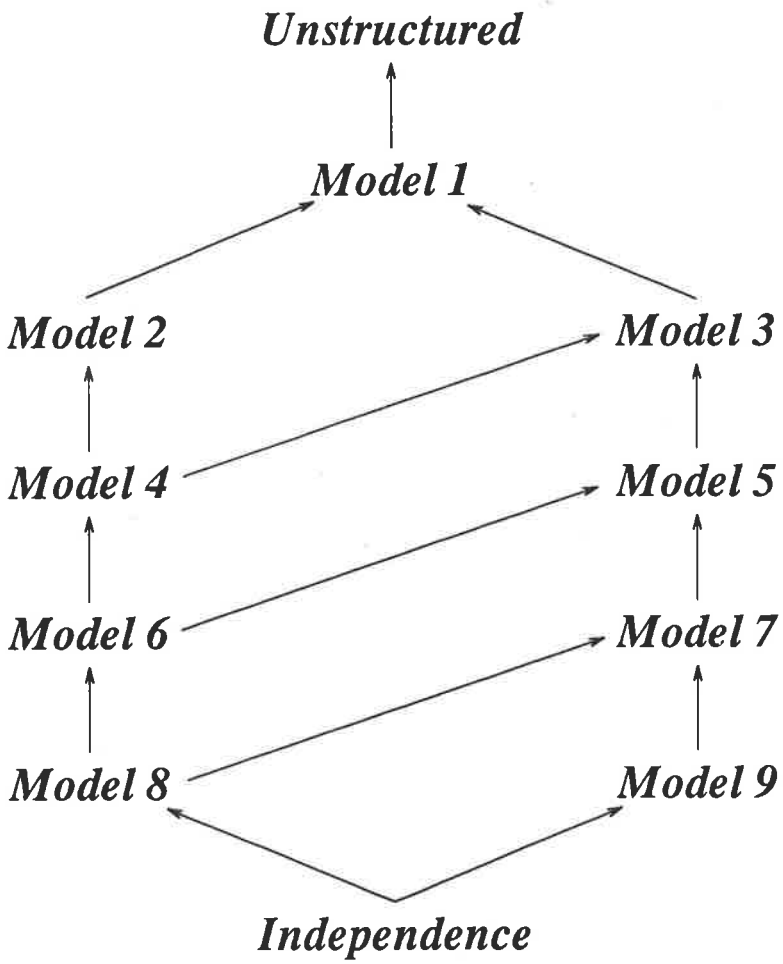

Figure 8.1: Correlation Model Lattice Diagram

Let the notation $\boldsymbol{R}_u(i:j,k:l)$ represent the submatrix of $\boldsymbol{R}_u$ indexed by rows i to j and columns k to l. The 6×6 matrix $\boldsymbol{\Sigma}_1$ of Model 1 is estimated by

$$\hat{\boldsymbol{\Sigma}}_1 = \{\boldsymbol{R}_u(1:6,1:6)+\boldsymbol{R}_u(7:12,7:12)+\boldsymbol{R}_u(13:18,13:18)+\boldsymbol{R}_u(19:24,19:24)\}/4.$$

Let

$$\boldsymbol{\Delta}_2 = \{\boldsymbol{R}_u(1:6,7:12) + \boldsymbol{R}_u(13:18,19:24)\}/2,$$

and

$$\boldsymbol{\Xi}_2 = \{\text{triu}(\boldsymbol{\Delta}_2) + \text{tril}(\boldsymbol{\Delta}_2)'\}/2,$$

where $\text{triu}(\boldsymbol{\Delta}_2)$ and $\text{tril}(\boldsymbol{\Delta}_2)$ are the upper and lower triangular (excluding the main diagonal elements) matrices of $\boldsymbol{\Delta}_2$ respectively. Then $\boldsymbol{\Sigma}_2$ is estimated by

$$\hat{\boldsymbol{\Sigma}}_2 = \text{diag}(\boldsymbol{\Delta}_2) + \boldsymbol{\Xi}_2 + \boldsymbol{\Xi}_2',$$

where $\text{diag}(\boldsymbol{\Delta}_2)$ is the matrix of diagonal elements of $\boldsymbol{\Delta}_2$. The submatrices $\boldsymbol{\Sigma}_3$ and $\boldsymbol{\Sigma}_4$ are estimated similarly. Estimation of the correlation parameters in Models 2-6 follow in a similar manner.

The above approach is not suitable for estimation of the correlation parameters in Models 7-9. This is because of the occurrence of products of correlation parameters. However suitable transformations of the individual matrices in the kronecker products offers an effective estimation technique.

The $q \times q$ equal correlation matrix,

$$\boldsymbol{R}_2 = \begin{bmatrix} 1 & \alpha & \alpha & \cdots & \alpha \\ \alpha & 1 & \alpha & \cdots & \alpha \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \alpha & \cdots & 1 \end{bmatrix} = \alpha \mathbf{1}_q \mathbf{1}_q' + (1-\alpha)\boldsymbol{I}_p,$$

has eigenvalues α and $(1-\alpha)$ of multiplicity 1 and $q-1$ respectively. Thus $\boldsymbol{R}_2$ can be diagonalized by $\boldsymbol{\Upsilon}_2'\boldsymbol{R}_2\boldsymbol{\Upsilon}_2$ to the diagonal matrix of the eigenvalues by the Spectral Decomposition Theorem. The column vectors of $\boldsymbol{\Upsilon}_2$ are the normalized eigenvectors of

$\boldsymbol{R}_2$. With $q = 3$, then

$$\boldsymbol{\Upsilon}_2 = \begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{2} & 1/\sqrt{6} \\ 1/\sqrt{3} & -1/\sqrt{2} & 1/\sqrt{6} \\ 1/\sqrt{3} & 0 & -2/\sqrt{6} \end{bmatrix}.$$

The third matrix in the kronecker product (8.5), $\boldsymbol{R}_3$ say, can be diagonalized to a matrix $\boldsymbol{\Upsilon}_3'\boldsymbol{R}_3\boldsymbol{\Upsilon}_3$ with diagonal elements $(1 + \rho_3)$ and $(1 - \rho_3)$, where

$$\boldsymbol{\Upsilon}_3 = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}.$$

The first matrix in (8.5), $\boldsymbol{R}_1$ say, can be diagonalized by $\boldsymbol{\Upsilon}_1'\boldsymbol{R}_1\boldsymbol{\Upsilon}_1$ to

$$\begin{bmatrix} (1 + \rho_2 + \rho_1 + \rho_4) & 0 & 0 & 0 \\ 0 & (1 - \rho_2 + \rho_1 - \rho_4) & 0 & 0 \\ 0 & 0 & (1 - \rho_1 - \rho_2 + \rho_4) & 0 \\ 0 & 0 & 0 & (1 + \rho_2 - \rho_1 - \rho_4) \end{bmatrix}$$

where

$$\boldsymbol{\Upsilon}_1 = 1/\sqrt{4} \begin{bmatrix} 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

If $\boldsymbol{\Upsilon} = \boldsymbol{\Upsilon}_1 \otimes \boldsymbol{\Upsilon}_2 \otimes \boldsymbol{\Upsilon}_3$, then by the properties of kronecker products,

$$\boldsymbol{\Upsilon}'\boldsymbol{R}\boldsymbol{\Upsilon} = (\boldsymbol{\Upsilon}_1'\boldsymbol{R}_1\boldsymbol{\Upsilon}_1) \otimes (\boldsymbol{\Upsilon}_2'\boldsymbol{R}_2\boldsymbol{\Upsilon}_2) \otimes (\boldsymbol{\Upsilon}_3'\boldsymbol{R}_3\boldsymbol{\Upsilon}_3), \tag{8.7}$$

which is a diagonal matrix. The diagonal elements of (8.7) are the products of the diagonal elements of individual matrices $\boldsymbol{\Upsilon}_j'\boldsymbol{R}_j\boldsymbol{\Upsilon}_j$, $j = 1, 2, 3$. For example the first diagonal element of (8.7) is $(1 + \rho_1 + \rho_2 + \rho_4)(2\alpha + 1)(1 + \rho_3)$ and the last is $(1 - \rho_1 + \rho_2 - \rho_4)(1 - \alpha + 1)(1 - \rho_3)$. The individual parameters are separated by various linear combinations of the diagonal elements of (8.7). The sum of the first 12 diagonal elements for example, equals $12(1 - \rho_1)$ and hence yields ρ_1.

Estimates of the correlation parameters are obtained by considering linear combinations of the diagonal elements of the transformed correlation matrix $\boldsymbol{\Upsilon}'\boldsymbol{R}_u\boldsymbol{\Upsilon}$. Estimates of the correlation parameters under Models 8 and 9 follow by similar transformations of $\boldsymbol{R}_u$.

8.6 Testing Correlation Models

It is difficult to test the validity of particular correlation structures under the GEE framework. The likelihood method is not available because of the semi-parametric nature of the GEE. However it is possible to use the asymptotic properties of the z-transforms to test the validity of Models 7-9.

Let $\boldsymbol{Y}_i$ be the 24×1 vector of responses for the i^{th} rat with mean vector $\boldsymbol{\mu}_i$. If $\boldsymbol{R}$ is the correct correlation matrix of the within-unit correlation, then the random vector $\boldsymbol{C}_i = \boldsymbol{A}_i^{-1/2}\boldsymbol{Y}_i/\phi$ has covariance matrix $\boldsymbol{R}$. The diagonal matrix $\boldsymbol{A}_i$ has diagonal elements $\mu_{ij}(1-\mu_{ij})$.

If Model 7 (or 8 or 9) is correct then the covariance matrix of the transform $\boldsymbol{\Upsilon}\boldsymbol{C}_i$ is (8.7). Thus the transform $\boldsymbol{\Upsilon}\boldsymbol{C}_i$ has zero correlations between the 24 observations.

Under the assumption of Normality, Hills (1969) [58] and Schweder and Spjøtvoll (1982) [127] used graphical techniques to study large sets of correlation coefficients. Moran (1980) [96] derived a formal testing procedure based on the distribution of the maximum correlation coefficient. An asymptotic formula for the tail of the distribution of the maximum of a set of product moment correlation coefficients was derived by Eagleson (1983) [35]. Under the Normality assumption, the procedure was shown to be quite accurate even for a small number of characteristics, p. It was also shown to be robust against failure of the Normality assumption. Cameron and Eagleson (1985) [17] expand the Poisson limit result used in Eagleson (1983) [35] and test the hypothesis that all the correlations between a set of variables are zero under Normality.

Brien *et al.* (1984) [16] study the large-sample joint distribution of the $\frac{1}{2}p(p-1)$

Fisher z-transforms,

$$Z_{jk} = \frac{1}{2}\log\left(\frac{1+r_{jk}}{1-r_{jk}}\right) = \tanh^{-1}(r_{jk}),$$

of the elements, r_{jk}, in a p variable sample correlation matrix. Let

$$\boldsymbol{Z} = (Z_{12},\ Z_{13},\ldots,\ Z_{1p},\ Z_{23},\ldots,\ Z_{(p-1)p})',$$

and $f = p(p-1)/2$. Define $\tilde{\boldsymbol{Z}}$ as $\sqrt{n}(\boldsymbol{Z}-\boldsymbol{\omega})$, where $\boldsymbol{\omega}$ is the vector of population correlations. The vector $\tilde{\boldsymbol{Z}}$ is asymptotically multivariate Normally distributed with zero mean vector and covariance matrix $\boldsymbol{\Lambda}$. The elements of $\boldsymbol{\Lambda}$ are complex functions of the population correlations (see Brien *et al.* (1984) [16]).

If $\boldsymbol{Z}$ were distributed as $N(\boldsymbol{\omega},\ \boldsymbol{\Lambda})$, then the quadratic form $Q = \boldsymbol{Z}'\boldsymbol{\Lambda}^{-1}\boldsymbol{Z}$ has the noncentral chi-squared distribution

$$\chi^2\left(\frac{1}{2}p(p-1),\ \boldsymbol{\omega}'\boldsymbol{\Lambda}^{-1}\boldsymbol{\omega}\right).$$

If the spectral decomposition of $\boldsymbol{\Lambda}$ is $\sum \kappa_i \boldsymbol{P}_i$, where $\boldsymbol{P}_i = \boldsymbol{P}_i'$, $\sum \boldsymbol{P}_i = \boldsymbol{I}_f$ and $\boldsymbol{P}_i\boldsymbol{P}_j = \delta_{ij}\boldsymbol{P}_i$ ($\delta_{ij} = 1$ if $i = j$, 0 otherwise), then $Q = \sum \kappa_i^{-1}\boldsymbol{Z}'\boldsymbol{P}_i\boldsymbol{Z}$. Each component of Q is independently distributed as noncentral chi-squared,

$$\kappa^{-1}\boldsymbol{Z}'\boldsymbol{P}_i\boldsymbol{Z} \sim \chi^2(\mathrm{tr}(\boldsymbol{P}_i),\ \kappa_i^{-1}\boldsymbol{\omega}'\boldsymbol{P}_i\boldsymbol{\omega})$$

where $\mathrm{tr}(\boldsymbol{P}_i)$ is the dimension of the invariant subspace onto which $\boldsymbol{P}_i$ projects.

The covariance matrix $\boldsymbol{\Lambda}$ has three projector matrices ($\boldsymbol{P}_0$, $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$) in its spectral decomposition under the null hypothesis of equal population correlations, $\rho_{jk} = \rho$, $\forall\ i,\ k$. The projector matrices define three mutually orthogonal invariant subspaces of the sample space. See Brien *et al.* (1984) [16] for further details including computational expressions of the quadratic components

$$Q_0 = \boldsymbol{Z}'\boldsymbol{P}_0\boldsymbol{Z}, \quad Q_1 = \boldsymbol{Z}'\boldsymbol{P}_1\boldsymbol{Z}, \quad Q_2 = \boldsymbol{Z}'\boldsymbol{P}_2\boldsymbol{Z}.$$

The following theorem holds,

Theorem 8.1 *If the correlations are all equal, the three quadratic forms* Q_i/κ_i, $i = 0, 1, 2$, *have independent chi-squared distributions given by,*

$$Q_0/\kappa_0 \sim \chi^2(1, \frac{1}{2}p(p-1)\omega^2/\kappa_0), \quad Q_1/\kappa_1 \sim \chi^2(p-1), \quad Q_2/\kappa_2 \sim \chi^2(\frac{1}{2}p(p-3)),$$

where ω *is the z-transform of the common correlation* ρ.

Proof: See Brien *et al.* (1984) [16].

The quadratic forms $\boldsymbol{Z}'\boldsymbol{P}_i\boldsymbol{Z}$ $(i = 0, 1, 2)$ can be calculated by fitting a notional linear model (see Ogawa and Ishii (1965) [101]) of the form

$$E(Z_{ij}) = \alpha_i + \alpha_j.$$

The 'main effects' sum of squares for testing $H_0 : \alpha_1 = \ldots = \alpha_p$ is Q_1 and the 'residual' or 'interaction' sum of squares is Q_2. The components are arranged in the analysis of variance Table 8.3. The overall test statistic for equality of correlations is the sum of the quadratic forms, $Q_1/\hat{\kappa}_1$ and $Q_2/\hat{\kappa}_2$.

Table 8.3:

Analysis of variance for equal correlations

Component	**Degrees of freedom**	**Sum of squares**	**Divisor**	χ^2
Grand mean	1	$Q_0 = \boldsymbol{Z}'\boldsymbol{P}_0\boldsymbol{Z}$	$\hat{\kappa}_0 = \frac{\{1+(p-1)\rho\}^2}{n(1+\rho)^2}$	$\frac{Q_0}{\hat{\kappa}_0}$
Main effects	$(p-1)$	$Q_1 = \boldsymbol{Z}'\boldsymbol{P}_1\boldsymbol{Z}$	$\hat{\kappa}_1 = \frac{p-(p-2)(1-\hat{\rho})^2}{2n(1+\hat{\rho})^2}$	$\frac{Q_1}{\hat{\kappa}_1}$
Interactions	$\frac{1}{2}p(p-3)$	$Q_2 = \boldsymbol{Z}'\boldsymbol{P}_2\boldsymbol{Z}$	$\hat{\kappa}_2 = \frac{1}{n(1+\hat{\rho})^2}$	$\frac{Q_2}{\hat{\kappa}_2}$
Equal Correlation	$\frac{1}{2}p(p-1)-1$			$\frac{Q_1}{\hat{\kappa}_1} + \frac{Q_2}{\hat{\kappa}_2}$
Total	$\frac{1}{2}p(p-1)$	$\boldsymbol{Z}'\boldsymbol{I}\boldsymbol{Z}$		

If $\hat{\omega}$ is the sample mean of the vector of z-transforms, $\boldsymbol{Z}$, then the inverse transform $\hat{\rho} = \tanh(\hat{\omega})$ is an obvious estimate of ρ. Under the assumption of equal correlations, then

$$\hat{\omega} \sim N(\omega, \ 2\kappa_0/(p(p-1))),$$

and specific hypotheses about the common correlation can be examined. Alternatively, the quadratic form, $Q_0/\hat{\kappa}_0$, can be evaluated at $\omega = 0$ giving a formal test of independence if the assumption of equal correlations is correct.

An equivalent approach is discussed in Seber (1984) [129] (pp. 98-101) which is a generalization of the method of Layard (1972) [74]. The random vectors $\boldsymbol{Y}_i$ were presumed to be i.i.d., with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{V}$. Let $\boldsymbol{U}$ be the $p \times 1$ vector of logarithms of sample variances, and $\boldsymbol{N} = \sqrt{n}(\boldsymbol{U},\ \boldsymbol{Z})'$, which has mean $\boldsymbol{\pi}$ and covariance matrix $\boldsymbol{\Gamma} = \boldsymbol{T}'\boldsymbol{\Omega}\boldsymbol{T}$. The vector $\boldsymbol{N}$ is asymptotically multivariate Normal by the multivariate central limit theorem. The matrix $\boldsymbol{\Omega}$ is the covariance matrix of sample variances and covariances and $\boldsymbol{T}_{uv} = \partial\boldsymbol{\pi}_v/\partial\boldsymbol{\sigma}_u$ where

$$\boldsymbol{\sigma}' = (\sigma_{11},\ \sigma_{22}, \ldots,\ \sigma_{pp},\ \sigma_{12}, \ldots,\ \sigma_{1p},\ \sigma_{2p}, \ldots,\ \sigma_{(p-1)p}),$$

that is the vector of variances and covariances. See Seber (1984) [129] (pp. 98-101) for a discussion of the estimation of the matrix $\boldsymbol{T}$.

Let the sample covariance matrix be $\boldsymbol{S} = \{s_{jk}\} = \sum_{i=1}^{n}(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})'/(n-1)$. It can be shown (e.g. Muirhead (1982) [98] (p. 42)) that asymptotically $\mathrm{cov}(\sqrt{n}s_{jk},\ \sqrt{n}s_{lm})$ is

$$E[(Y_j - \mu_j)(Y_k - \mu_k)(Y_l - \mu_l)(Y_m - \mu_m)] - E[(Y_j - \mu_j)(Y_k - \mu_k)]E[(Y_l - \mu_l)(Y_m - \mu_m)]$$

which can be estimated by the moment estimator

$$\begin{aligned}&\frac{1}{n-1}\sum_{i=1}^{n}[(Y_{ij} - \bar{Y}_j)(Y_{ik} - \bar{Y}_k)(Y_{il} - \bar{Y}_l)(Y_{im} - \bar{Y}_m)] - \\ &\frac{1}{n-1}\sum_{i=1}^{n}[(Y_{ij} - \bar{Y}_j)(Y_{ik} - \bar{Y}_k)]\frac{1}{n-1}\sum_{i=1}^{n}[(Y_{il} - \bar{Y}_l)(Y_{im} - \bar{Y}_m)]. \qquad (8.8)\end{aligned}$$

This leads to an estimator of $\boldsymbol{\Omega}$. Replace $\bar{\boldsymbol{Y}}$ and the divisor $n-1$ by $\hat{\boldsymbol{\mu}}_i$ and $n-q$ respectively, when $E(\boldsymbol{Y}_i) = \boldsymbol{\mu}_i$ (i.e. some design is assumed), q are the number of parameters in the mean model.

Test of hypothesis about the covariance matrix of the response matrix $\boldsymbol{Y}$ takes the form $H_0 : \boldsymbol{G}\boldsymbol{\pi} = \boldsymbol{0}$, where $\boldsymbol{G}$ is a $g \times k$ matrix with full column rank, and $k = p + \frac{1}{2}p(p-1)$.

When H_0 is true, then $\boldsymbol{GN}$ is approximately $N(\mathbf{0},\ \boldsymbol{G\Gamma G'})$ and

$$(\boldsymbol{GN})'(\boldsymbol{G\Gamma G'})^{-1}\boldsymbol{GN} \sim \chi^2_g.$$

This approach is suitable for examining Models 7-9 as well as Models 1-6, although $\boldsymbol{G}$ will generally be quite complicated.

8.7 A Mixed Effects Model

Zeger, Albert and Liang (1988) [168] considered mixed effects for generalized linear models (GLM's) in the analysis of longitudinal data. The mean regression parameters are estimated by GEE.

Let $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ be $m \times 1$ and $q \times 1$ vectors of explanatory variables respectively at observation j for subject i. Suppose $\boldsymbol{b}_i$ is a $q \times 1$ vector of random effects with covariance matrix $\boldsymbol{\Psi}$. The mixed effects GLM is defined as

$$h(u_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i, \quad \text{var}(y_{ij}) = \phi g(\mu_{ij}), \tag{8.9}$$

where $u_{ij} = E(y_{ij}|\boldsymbol{b}_i)$, and $\boldsymbol{b}_i$ is an independent variable with mixture distribution F. This is termed a subject specific (SS) model in Zeger, Albert and Liang (1988) [168]. The marginal mean is

$$\mu_{ij} = E(y_{ij}) = E[E(y_{ij}|\boldsymbol{b}_i)].$$

Suppose the mixture distribution F is the Normal distribution with mean $\mathbf{0}$. This choice of F simplifies the marginal mean. Zeger, Albert and Liang (1988) [168] give exact expressions for the log and probit link for this choice of F. No exact closed-form expression exists for the marginal mean using a logit link. Instead Zeger, Albert and Liang (1988) [168] approximate the logistic function by a cumulative Gaussian approximation yielding $\text{logit}(\mu_{ij}) \simeq a_l(\boldsymbol{\Psi})\ .\ \boldsymbol{x}'_{ij}\boldsymbol{\beta}$, where $a_l(\boldsymbol{\Psi}) = |c^2\boldsymbol{\Psi}\boldsymbol{z}_{ij}\boldsymbol{z}'_{ij} + \mathbf{I}|^{-q/2}$ and $c = 16\sqrt{3}/(15\pi)$. The parameter vector $\boldsymbol{\beta}$ is estimated by the GEE (5.8). An approximation to the marginal covariance for the i^{th} subject is

$$\boldsymbol{V}_i = \text{cov}(\boldsymbol{Y}_i) \simeq \boldsymbol{L}_i\boldsymbol{Z}_i\boldsymbol{\Psi}\boldsymbol{Z}'_i\boldsymbol{L}_i + \phi\boldsymbol{A}_i = \tilde{\boldsymbol{V}}_i, \tag{8.10}$$

where $\boldsymbol{L}_i = \text{diag}\{\partial h^{-1}(u)/\partial u,\ u = \boldsymbol{x}'_{ij}\boldsymbol{\beta},\ j = 1, \ldots,\ p\}$ and $\boldsymbol{A}_i = \text{diag}\{g(\mu_{ij}),\ j = 1, \ldots,\ p\}$. A crude estimate of the random effects covariance matrix $\boldsymbol{\Psi}$ is the moment estimator,

$$\hat{\boldsymbol{\Psi}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{Z}_i'\boldsymbol{Z}_i)^{-1}\boldsymbol{Z}_i'\hat{\boldsymbol{L}}_i^{-1}[(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)' - \hat{\phi}\hat{\boldsymbol{A}}_i]\hat{\boldsymbol{L}}_i^{-1}\boldsymbol{Z}_i(\boldsymbol{Z}_i'\boldsymbol{Z}_i)^{-1}. \tag{8.11}$$

Note the diagonal elements of the approximation (8.10) are

$$E(y_{ij} - \mu_{ij})^2 \simeq \phi g(\mu_{ij})(\hat{\boldsymbol{L}}_{i_{jj}})^2\boldsymbol{z}'_{ij}\hat{\boldsymbol{\Psi}}\boldsymbol{z}_{ij},\ j = 1, \ldots,\ p. \tag{8.12}$$

The moment estimator of the dispersion parameter using (8.12) is

$$np\hat{\phi} = \sum_{i=1}^{n}\sum_{j=1}^{p}\frac{(y_{ij} - \hat{\mu}_{ij})^2 - (\hat{\boldsymbol{L}}_{i_{jj}})^2\boldsymbol{z}'_{ij}\hat{\boldsymbol{\Psi}}\boldsymbol{z}_{ij}}{g(\hat{\mu}_{ij})}. \tag{8.13}$$

To model the dispersion parameters in a similar manner as the marginal case considered in §6.3, I suggest using a second order GEE for ϕ with the response variable

$$d_{ij} = \frac{(y_{ij} - \mu_{ij})^2 - (\boldsymbol{L}_{i_{jj}})^2\boldsymbol{z}'_{ij}\boldsymbol{\Psi}\boldsymbol{z}_{ij}}{g(\mu_{ij})}.$$

8.8 Analysis of the Rat Teeth Data

The estimated parameters in mean model (8.1) were tested and non-significant terms were progressively dropped. All two-way interaction terms containing treatment B were non-significant. The interaction of C and TB is also non-significant. Neither is the main effect of treatment B statistically significant. The above applies equally to correlation Models 1-9 as well as the independence and fully parameterized or unstructured cases. Parameter estimates for the reduced mean model are displayed in Table 8.4. Models 4 and 6 are excluded because they are similar to Models 3 and 5 respectively, except the latter models have slightly lower AIC values. Models 8 and 9 are also excluded because they are similar, not unexpectedly, to Model 7.

The parameter estimates over all the parametric models are generally similar. The standard errors are lowest for the unstructured case and increase as the number of

Table 8.4:

Param.	Model						
	I	U	1	2	3	5	7
β_0	-1.220	-1.142	-1.070	-1.094	-1.296	-1.163	-1.192
	(.325)	(.320)	(.343)	(.349)	(.360)	(.348)	(.336)
CC	3.258	2.381	2.793	2.816	3.081	3.188	3.220
	(.414)	(.348)	(.401)	(.406)	(.425)	(.431)	(.420)
A1	-0.101	-0.031	-0.089	-0.109	-0.103	-0.048	-0.074
	(.936)	(.474)	(.701)	(.704)	(.783)	(.927)	(.925)
A2	-0.612	-0.232	-0.434	-0.464	-0.584	-0.620	-0.605
	(.551)	(.384)	(.495)	(.497)	(.504)	(.546)	(.547)
A3	-0.969	-0.687	-0.893	-0.933	-0.982	-0.956	-0.958
	(.588)	(.407)	(.507)	(.504)	(.513)	(.587)	(.586)
TB	0.905	0.538	0.718	0.755	1.107	0.893	0.898
	(.213)	(.119)	(.168)	(.170)	(.223)	(.223)	(.219)
MOL2	0.064	-0.063	0.016	0.042	0.201	0.045	0.028
	(.285)	(.109)	(.207)	(.204)	(.241)	(.261)	(.261)
MOL3	-0.056	0.262	-0.038	-0.032	0.306	0.033	0.037
	(.318)	(.207)	(.262)	(.265)	(.270)	(.310)	(.313)
A1xTB	-0.475	-0.455	-0.491	-0.482	-0.545	-0.522	-.504
	(.395)	(.248)	(.319)	(.335)	(.423)	(.422)	(.410)
A2xTB	-0.647	-0.575	-0.711	-0.676	-0.627	-0.660	-0.655
	(.389)	(.288)	(.347)	(.353)	(.394)	(.396)	(.391)
A3xTB	-1.257	-0.714	-1.114	-1.064	-1.106	-1.250	-1.255
	(.436)	(.242)	(.324)	(.311)	(.362)	(.420)	(.422)
CCxMOL2	-1.428	-0.375	-1.017	-1.015	-1.280	-1.401	-1.383
	(.361)	(.140)	(.266)	(.263)	(.295)	(.331)	(.333)
A1xMOL2	0.208	0.286	0.290	0.300	0.265	0.200	0.207
	(.734)	(.204)	(.503)	(.500)	(.577)	(.722)	(.717)
A2xMOL2	0.784	0.405	0.627	0.637	0.721	0.766	0.763
	(.409)	(.141)	(.317)	(.315)	(.319)	(.395)	(.397)
A3xMOL2	0.332	0.038	0.237	0.261	0.322	0.305	0.302
	(.510)	(.171)	(0.357)	(.345)	(.375)	(.461)	(.456)
CCxMOL3	-0.914	-0.357	-0.426	-0.447	-0.951	-.998	-1.002
	(.369)	(.230)	(.293)	(.296)	(.310)	(.397)	(.366)
A1xMOL3	0.870	0.836	0.908	0.916	0.944	0.867	0.870
	(.548)	(.228)	(.321)	(.319)	(.388)	(.539)	(.535)
A2xMOL3	0.282	0.017	0.133	0.151	0.243	0.248	0.244
	(.414)	(.230)	(.338)	(.332)	(.319)	(.397)	(0.398)
A3xMOL3	0.219	0.202	0.152	0.185	0.355	0.225	0.229
	(.392)	(.236)	(.272)	(.267)	(.291)	(.385)	(.390)
TBxMOL2	-2.492	-1.673	-2.134	-2.219	-2.752	-2.479	-2.486
	(.247)	(.158)	(.205)	(.206)	(.251)	(.245)	(.246)
TBxMOL3	-3.815	-3.482	-3.734	-3.715	-4.034	-3.797	-3.801
	(.290)	(.216)	(.265)	(.265)	(.291)	(.292)	(.291)
ϕ	0.985	0.841	0.921	0.916	0.954	0.980	0.983
	(.183)	(.082)	(.221)	(.483)	(.428)	(.199)	(.187)
log (G)	66.5	85.0	71.2	71.6	69.4	66.6	67.0
AIC	1497	2199	1953	1934	1802	1811	1725

Table 8.5:

Analysis of variance for equal correlations

Component	Degrees of freedom	χ^2
Grand mean	1	8.046
Main effects	23	70.387
Interactions	252	801.118
Equal Correlation	275	871.505
Total	276	

correlation parameters are reduced. This is because the smaller the number of correlation parameters, the further the parametric correlation model is from the observed correlation pattern, especially for large correlation matrices. The statistics $\log(G)$ and AIC also reflect this result.

The standard error of ϕ under Models 2 and 3 (and Model 4) is quite high relative to the other correlation models. In Models 3 and 4, applying (8.4) to $\boldsymbol{\Sigma}_2$ and $\boldsymbol{\Sigma}_3$ (and to $\boldsymbol{\Sigma}_4$ for Model 3) and not to $\boldsymbol{\Sigma}_1$ may be a contributing factor.

Except for the high standard error of ϕ, Model 2 appears to be almost as good as Model 1. The standard errors of the estimates of the mean parameters are quite similar. This is reflected in the $\log(G)$ statistic (Model 2 is actually marginally higher than Model 1). Similarly the AIC statistic acknowledges moving from Model 1 to Model 2 does not incur a big penalty especially compared to Models 3-9 (though the reduction in the number of correlation parameters is more pronounced for these models).

Testing the kronecker product correlation matrices of Models 7-9 was discussed in §8.6 using the method proposed by Brien *et al.* (1984) [16]. Table 8.5 displays the analysis of variance for equal correlation of the transform $\boldsymbol{\Upsilon}\boldsymbol{C}_i$ (recall $\boldsymbol{C}_i = \boldsymbol{A}_i^{-1/2}\boldsymbol{Y}_i/\phi$) under Model 7. Similar values occur for Models 8 and 9.

The magnitude of the test statistic, 871.5, convincingly rejects the hypothesis of

equal correlation, and consequently testing that the common correlation is zero is not appropriate.

The equivalent method discussed by Seber (1984) [129] (pp. 98-101) and introduced in §8.6 was also used for Models 7 to 9. To test the hypothesis of zero correlations for $\boldsymbol{\Upsilon C}_i$, then $\boldsymbol{G} = [\mathbf{0}_p \ \boldsymbol{I}_{p(p-1)/2}]$ in the hypothesis $H_0 : \ \boldsymbol{G\pi} = \mathbf{0}$. Unfortunately the estimate of the 300×300 covariance matrix $\boldsymbol{\Gamma} = \boldsymbol{T}'\boldsymbol{\Omega}\boldsymbol{T}$ is highly singular (rank 103). This is because the matrix of moment estimators (8.8) (i. e. the estimate of $\boldsymbol{\Omega}$) is highly singular (rank 103). The discrete nature of the data is probably responsible for the singularity. Using a pseudo-inverse instead of the inverse $(G\Gamma G)^{-1}$ with such a severe singularity, the test statistic must be viewed with extreme caution. The test statistic for Model 7 for example, is 5.9850×10^4 (compared to 871.5 in Table 8.5).

In Figure 8.2a the histogram of the correlations of the transformation $\boldsymbol{\Upsilon C}_i$ is displayed. The correlations are dependent and correlation values need not be large to be significant given the large sample size. The contributions in the tails of the histogram in Figure 8.2a is large enough to reject the hypothesis of zero correlations.

Given that the second test procedure (Seber (1984) [129] (pp. 98-101)) discussed in §8.6 failed for this example, histograms such as Figure 8.2b were used to examine Models 1 to 6. The histogram in Figure 8.2b displays the distribution of the differences between the 276 correlation parameter estimates for the unstructured case and the corresponding correlations under correlation Model 1. The differences are slightly asymmetric, with mean and mode both nearly zero. Though not much emphasis should be placed on Figure 8.2b, because the estimates of the parameters of the Model 1 were derived from the unstructured parameter estimates, the magnitude of the differences are generally not very big, and given the size of the correlation matrix then Model 1 may be a reasonable correlation model. Obviously histograms under Models 2-6 (not shown) become worse as the number of correlation parameters decreases but, particularly for Model 2, generally appear to be adequate.

The correlation assumptions under Model 1 are minimal compared to Models 2-9

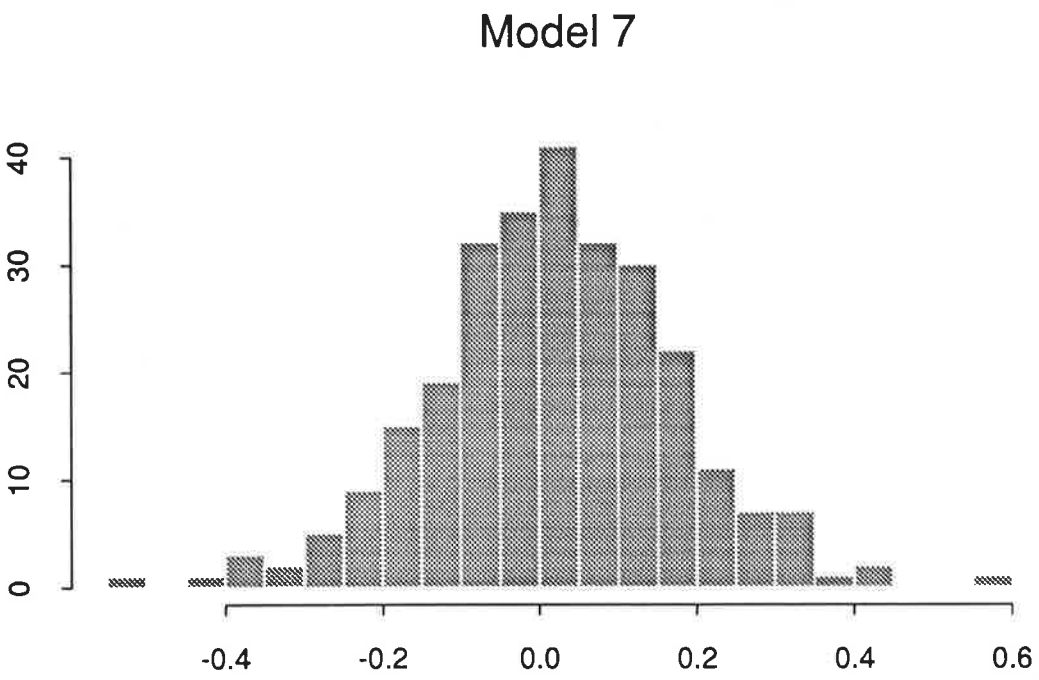

Figure 8.2a

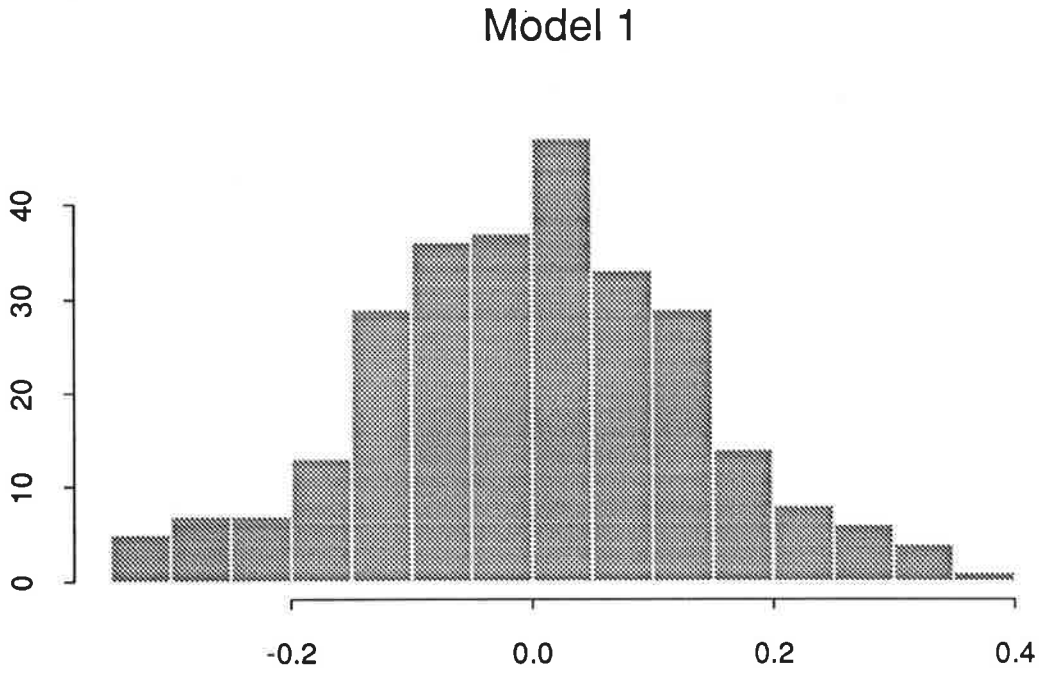

Figure 8.2b

and the Independence case. Except for Model 2, AIC (and to a lesser extent $\log(G)$) indicates the poorer fitting of the other parametric models compared to Model 1. This is reflected in the higher standard errors for Models 3-9. Even under Model 1 a substantial reduction in the number of correlation parameters over the unstructured case is achieved. Consequently I consider Model 1 as the preferred parametric model. Model 2 is handicapped by the high standard error of ϕ. However as stated earlier the estimates of the mean parameters and their standard errors are close to those under Model 1 (reflected in AIC and $\log(G)$). As well there is a further and significant reduction in the number of correlation parameters. Thus Model 2 can be considered to be almost as good as Model 1 in Table 8.4.

Now suppose mean model (8.1) is expanded to include random effects specified by left-right and within-molar indicators, that is a mixed effects model. The random effects are assumed to have zero mean. This mixed effects model follows from the assumption that there are no biological differences between the left and right jaw (symmetry) and within a molar. The results of §8.7 can be applied. The estimation procedure using (5.8), (8.11) and (8.13) failed to converge. The estimator (8.11) became unstable very quickly. Zeger, Albert and Liang (1988) [168] reported this algorithm also failed to converge for the one example they analysed because there was little information about $\boldsymbol{\Psi}$ (a scalar there) in the data they examined. They examined the regression coefficients over a range of values for $\boldsymbol{\Psi}$.

However an estimate of $\boldsymbol{\Psi}$ was obtained by first making $\boldsymbol{\Psi}$ fixed at some initial starting value (e. g. $\hat{\boldsymbol{\Psi}}_0 = \boldsymbol{I}_2$), and (5.8) and (8.13) were iterated until convergence. A new estimate of $\boldsymbol{\Psi}$, $\hat{\boldsymbol{\Psi}}_1$ say, was obtained by (8.11) with the current estimates of $\boldsymbol{\beta}$ and ϕ. New estimates of $\boldsymbol{\beta}$ and ϕ are obtained for $\boldsymbol{\Psi}$ fixed at $\hat{\boldsymbol{\Psi}}_1$ until convergence. This procedure was continued until convergence of the estimate of $\boldsymbol{\Psi}$ was achieved. The final estimate of $\boldsymbol{\Psi}$ was

$$\begin{bmatrix} 0.960 & -0.019 \\ -0.019 & 0.379 \end{bmatrix}.$$

Estimates of $\boldsymbol{\beta}$ and ϕ for two other fixed choices of $\boldsymbol{\Psi}$, that is

$$\begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 0.5 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0.3 & 0.25 \\ 0.25 & 0.05 \end{bmatrix},$$

were calculated. The second of these matrices was chosen by tracking the moment estimate (8.11) before the instability of the estimating procedure of §8.7 occurred. The estimates are given in Table 8.6. A number of features are worthy of comment. The final mean parametric model conditional on the random effects is the same as the final unconditional mean parametric model (8.1). It is interesting to note, with respect to the goodness of fit statistic $\log(G)$, the worst model corresponds to the third model in Table 8.6 (i.e. the moment estimate of $\boldsymbol{\Psi}$). With the same number of parameters, the second model in Table 8.6 is superior by this criterion. The first model is roughly on par with the third model. The second model in Table 8.6 is also the best using AIC and the third model is marginally ahead of the first. With respect to the various correlation models displayed in Table 8.4, the mixed effects model does not perform as well. There is further structure in the variance that has not been accounted for under the mixed effects model.

Now suppose the linear predictor (8.1) is expanded to incorporate a Left-Right (LR) fixed effect and a Within Molar (WM) fixed effect, that is (8.2). Correlation Models 1, 5 and 9 were considered as well as the independence and unstructured cases. The final parametric form of (8.2) incorporates all the significant terms of the final form (8.1). Not surprisingly the interaction of Molars and Within Molars is significant. What is surprising is that the Left-Right effect is significant. The estimates are displayed in Table 8.7. Model 4 was considered as well, and has estimate $\hat{\phi} = 1.1043$ with standard error 0.168. This standard error differs little from the corresponding standard error in the other models displayed in Table 8.7. This is in contrast to Model 4 with mean model (8.1).

As expected, the Unstructured model has the highest goodness of fit values. What is interesting, is that the Independence model appears to be equally as good as Models

Table 8.6:

Param.	Matrix		
	$\Psi_{11} = 1.0$ $\Psi_{12} = 0.0$ $\Psi_{22} = 0.500$	$\Psi_{11} = 0.3$ $\Psi_{12} = 0.250$ $\Psi_{22} = 0.050$	$\Psi_{11} = 0.960$ $\Psi_{12} = -0.019$ $\Psi_{22} = 0.379$
$\hat{\beta}_0$	-1.293 (.402)	-1.235 (.366)	-1.310 (.406)
CC	3.560 (.513)	3.348 (.464)	3.629 (.519)
A1	0.067 (1.137)	0.028 (1.036)	0.058 (1.162)
A2	-0.776 (.626)	-0.684 (.581)	-0.797 (.639)
A3	-1.094 (.659)	-1.010 (.615)	-1.100 (.675)
TB	1.058 (.244)	0.981 (.225)	1.069 (.247)
MOL2	0.111 (.323)	0.111 (.300)	0.117 (.323)
MOL3	0.004 (.317)	-0.030 (.310)	0.0002 (.322)
A1xTB	-0.569 (.415)	-0.524 (.431)	-0.577 (.480)
A2xTB	-0.761 (.468)	-0.709 (.423)	-0.770 (.473)
A3xTB	-1.542 (.530)	-1.397 (.483)	-1.566 (.535)
CCxMOL2	-1.744 (.415)	-1.584 (.381)	-1.758 (.418)
A1xMOL2	0.235 (.694)	0.212 (.824)	0.247 (.936)
A2xMOL2	0.929 (.485)	0.848 (.441)	0.944 (.491)
A3xMOL2	0.417 (.605)	0.360 (.552)	0.423 (.612)
CCxMOL3	-1.228 (.388)	-1.063 (.368)	-1.248 (.395)
A1xMOL3	0.991 (.694)	0.922 (.623)	1.003 (.712)
A2xMOL3	0.424 (.485)	0.351 (.441)	0.430 (.493)
A3xMOL3	0.259 (.458)	0.240 (.421)	0.265 (.469)
TBxMOL2	-2.903 (.296)	-2.701 (.269)	-2.949 (.299)
TBxMOL3	-4.405 (.347)	-4.083 (.317)	-4.441 (.350)
ϕ	0.939	0.958	0.938
$\log(G)$	50.1	53.6	49.5
AIC	1487	1509	1489

Table 8.7:

Param.	Model									
	I		U		1		5		9	
$\hat{\beta}_0$	-1.417	(.344)	-1.432	(.347)	-1.307	(.363)	-1.460	(.370)	-1.436	(.355)
CC	3.291	(.417)	2.841	(.368)	3.136	(.434)	3.235	(.436)	3.184	(.415)
A1	-0.095	(.945)	-0.132	(.538)	-0.161	(.779)	-0.087	(.840)	-0.092	(.836)
A2	-0.608	(.558)	-0.382	(.402)	-0.557	(.524)	-0.658	(.513)	-0.571	(.415)
A3	-0.974	(.595)	-0.838	(.450)	-0.950	(.569)	-0.861	(.596)	-0.795	(.605)
TB	0.917	(.215)	0.531	(.141)	0.820	(.201)	0.850	(.225)	0.844	(.219)
LR	0.453	(.145)	0.445	(.126)	0.454	(.141)	0.454	(.144)	0.467	(.145)
MOL2	0.475	(.294)	-0.040	(.116)	0.048	(.215)	0.365	(.241)	0.332	(.222)
MOL3	-0.162	(.331)	-0.008	(.241)	-0.348	(.318)	-0.084	(.340)	-0.207	(.344)
WMOL	-0.090	(.115)	-0.068	(.078)	-0.093	(.110)	-0.060	(.113)	-0.061	(.114)
A1xTB	-0.493	(.408)	-0.461	(.254)	-0.485	(.338)	-0.501	(.426)	-0.460	(.425)
A2xTB	-0.668	(.401)	-0.657	(.306)	-0.799	(.377)	-0.771	(.389)	-0.777	(.388)
A3xTB	-1.273	(.441)	-0.910	(.276)	-1.276	(.372)	-1.356	(.403)	-1.412	(.418)
CCxMOL2	-1.348	(.374)	-0.360	(.144)	-0.978	(.290)	-1.123	(.309)	-1.190	(.298)
A1xMOL2	0.206	(.738)	0.369	(.255)	0.360	(.600)	0.166	(.680)	0.173	(.700)
A2xMOL2	0.785	(.417)	0.480	(.168)	0.688	(.353)	0.683	(.395)	0.664	(.412)
A3xMOL2	0.290	(.522)	0.011	(.217)	0.294	(.355)	0.125	(.452)	0.144	(.477)
CCxMOL3	-0.921	(.370)	-0.643	(.242)	-0.644	(.358)	-0.972	(.371)	-0.952	(.370)
A1xMOL3	0.878	(.548)	0.893	(.218)	0.995	(.416)	0.873	(.482)	0.842	(.492)
A2xMOL3	0.280	(.418)	0.133	(.234)	0.238	(.402)	0.198	(.377)	0.178	(.381)
A3xMOL3	0.218	(.396)	0.270	(.234)	0.218	(.326)	0.175	(.390)	0.153	(.401)
TBxMOL2	-2.617	(.260)	-2.138	(.194)	-2.465	(.255)	-2.648	(.264)	-2.708	(.272)
TBxMOL3	-3.854	(.291)	-3.377	(.230)	-3.996	(.306)	-3.881	(.302)	-3.946	(.303)
MOL2xWM	-0.918	(.191)	-0.636	(.124)	-0.947	(.184)	-1.033	(.193)	-0.970	(.202)
MOL3xWM	0.214	(.186)	0.159	(.155)	0.181	(.191)	0.204	(.183)	0.353	(.183)
ϕ	1.004	(.992)	0.909	(.128)	1.046	(.167)	1.053	(.237)	1.089	(.168)
$\log(G)$	83.2		100.0		83.6		82.7		81.9	
AIC	1531		2218		1958		1815		1695	

1 and 5 by the $\log(G)$ statistic. This is definitely not the case in Table 8.4. Model 9 also fares well compared to these models. Adding key significant terms to the model (8.1) has improved the goodness of fit criterion $\log(G)$ for Models 5 and 9 (and the Independence Model) relative to Model 1 (compared with Table 8.4). The AIC statistic on the other hand indicates the poor fitting qualities of the Independence model as would be expected here (and reflected in the higher standard errors). Similarly the magnitude of the differences of the model AIC statistics has changed little when moving from the mean model (8.1) to the extended mean model (8.2). The AIC statistic provides a clearer indication of the better fitting parametric models. As demonstrated in chapter 7, $\log(G)$ must be used with some caution. The preferred model in Table 8.7 is again Model 1. Another unusual feature of the Independence model is the high standard error for ϕ. Both $\log(G)$ and to a lesser extent AIC show the improvement in goodness of fit from moving from (8.1) to (8.2).

A mixed-effects model, with random effects LR and WM, was applied to (8.2). This model failed to converge except when the random effects covariance matrix was fixed indefinitely.

Table 8.8 displays the correlation parameter estimates for the kronecker product correlation Models 7-9 under (8.1) and Model 9 under (8.2).

Figure 8.3 is a plot of the predicted probability of severe decay against the within-row molar position under the final form of (8.1) with correlation Model 1. The index for the molar axis corresponds to molar/(within-molar) position within a row, e. g. indices 1 and 2 refer to the first and second within-molar site of the first molar. The solid line is the Cariogenic control group (Diet 2), the broken line is the Non-Cariogenic control group (Diet 1) and the dotted line is either Diet 3, 4 or 5. Note the probabilities under Diets 6, 7 and 8 are the same as Diet 2 because of the non-significance of Treatment B. The top row may be either the left or right top row. Similarly for the bottom row.

A number of key features become evident from Figure 8.3. The higher the amount of Treatment A applied, the lower the probability of severe decay for both top and bottom

Table 8.8:

Param.	Model			
	7	8	9	9*
α	0.2463	0.2463	0.2463	0.2527
ρ_1	0.2310	0.2310	0.2309	0.2278
ρ_2	0.2197	0.1694	0.2196	0.2236
ρ_3	0.5046	0.5047	0.5045	0.5257
ρ_4	0.1190			

9* is correlation Model 9 under the extended linear predictor (8.2)

rows. There are fundamental differences in the underlying probabilistic mechanism between the top and bottom rows. One possible explanation of the indicated probability structure is that the top row is really the bottom row and vice versa. This follows if it is assumed the chance of decay is greater on the bottom than on the top. Further, if decay increases towards the rear of the mouth, then index 1 is now the second within-molar site of the last molar.

Figure 8.4 is a plot of the observed probabilities for the top-left row for Diets 1 to 8. As in Figure 8.3, Diets 1 and 2 are the reference diets. Figures 8.5, 8.6 and 8.7 are the observed probabilities for the top-right, bottom-left and bottom-right rows respectively.

The cavity reducing effect of Treatment A is fairly obvious as is the ineffectiveness of Treatment B. The similarity of the bottom-left and bottom-right probability pattern supports the original assumption that these measurements belong to the same jaw. The significance of the Left-Right effect in the extended mean model (8.2) is understandable given the obvious graphical differences between the top-left and top-right observed probability structures. However there is still sufficient likeness between the top-left and top-right patterns (compared to the bottom patterns) to continue advocating they be-

long to the same jaw. Figures 8.8 and 8.9 are plots of the expected probabilities of the extended mean model (8.2) under correlation Model 1. They indicate that the probability of severe decay is shifted downwards for the right-hand side of the mouth. The extended mean model (8.2) tends to be better than the final form of (8.1) at capturing the underlying process that is indicated in Figures 8.4-8.7.

8.9 Discussion

The difficulty of analysing complex non-Normal data containing structure which indicates possible parametric covariance matrices is a common problem. The GEE approach can provide a natural and usually straightforward method, reminiscent of the regression techniques employed in Normal theory. The parametric covariance matrices are neatly and explicitly incorporated into the estimating procedure.

Modelling the covariance structure such as for the rat teeth data can lead to improvements in the efficiency of the estimating procedure and possibly lower standard errors. A principal goal of the researcher is to uncover and model influential factors or explanatory variables of the mean of the response variable, and though usually of secondary importance, to adequately account for the underlying covariance structure. In modelling the covariance structure, the unobserved error process may be examined. The choice of the covariance structure can be critical to the actual mean model selected. It may also affect the stability of the estimating procedure and inflate the estimates of the standard errors of the parameters, as was seen in chapter 7. Accordingly, it is worthwhile devoting effort to try to determine the suitability of various covariance structures for a given data set.

As was discussed, asymptotic tests on the parametric covariance structures can be developed and sometimes the parametric structure lends itself to possible specialized tests (such as for Models 7-9). Despite the difficulties encountered with the rat teeth data in applying the more general tests to different parametric covariance structures,

the techniques discussed in this chapter can be implemented successfully to other data sets.

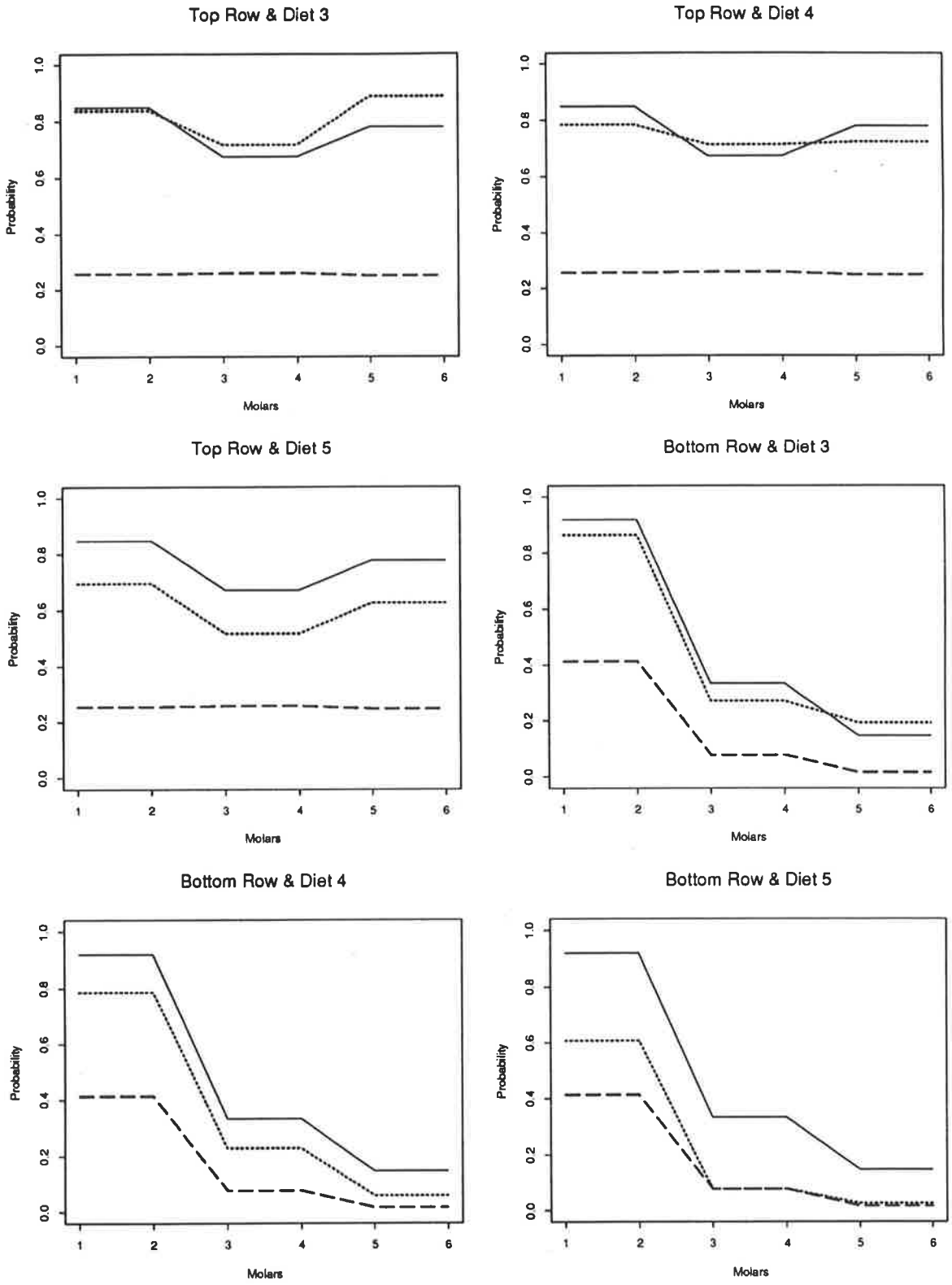

Figure 8.3: Correlation Model 1

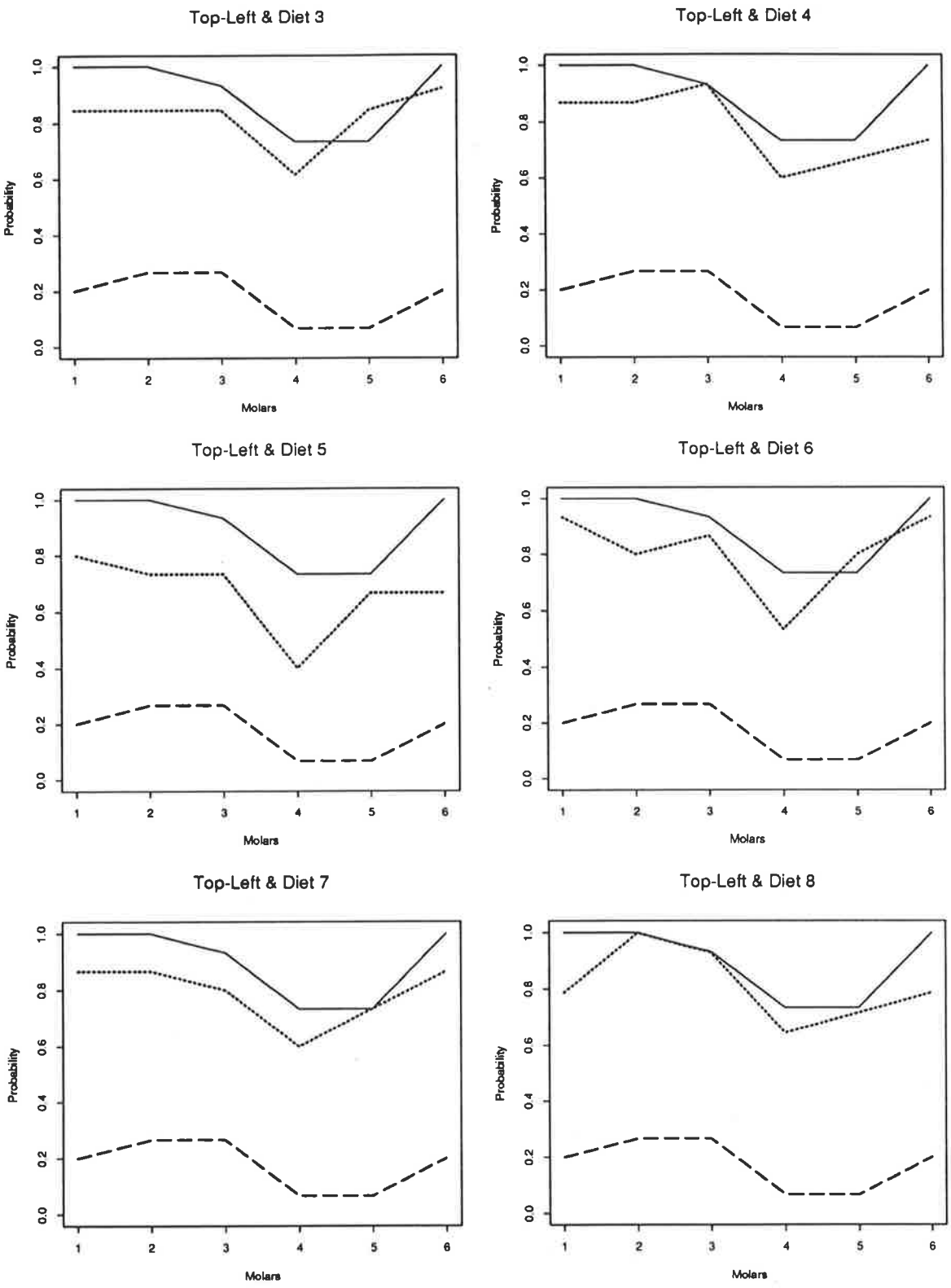

Figure 8.4: Observed probabilities for top-left portion of the mouth

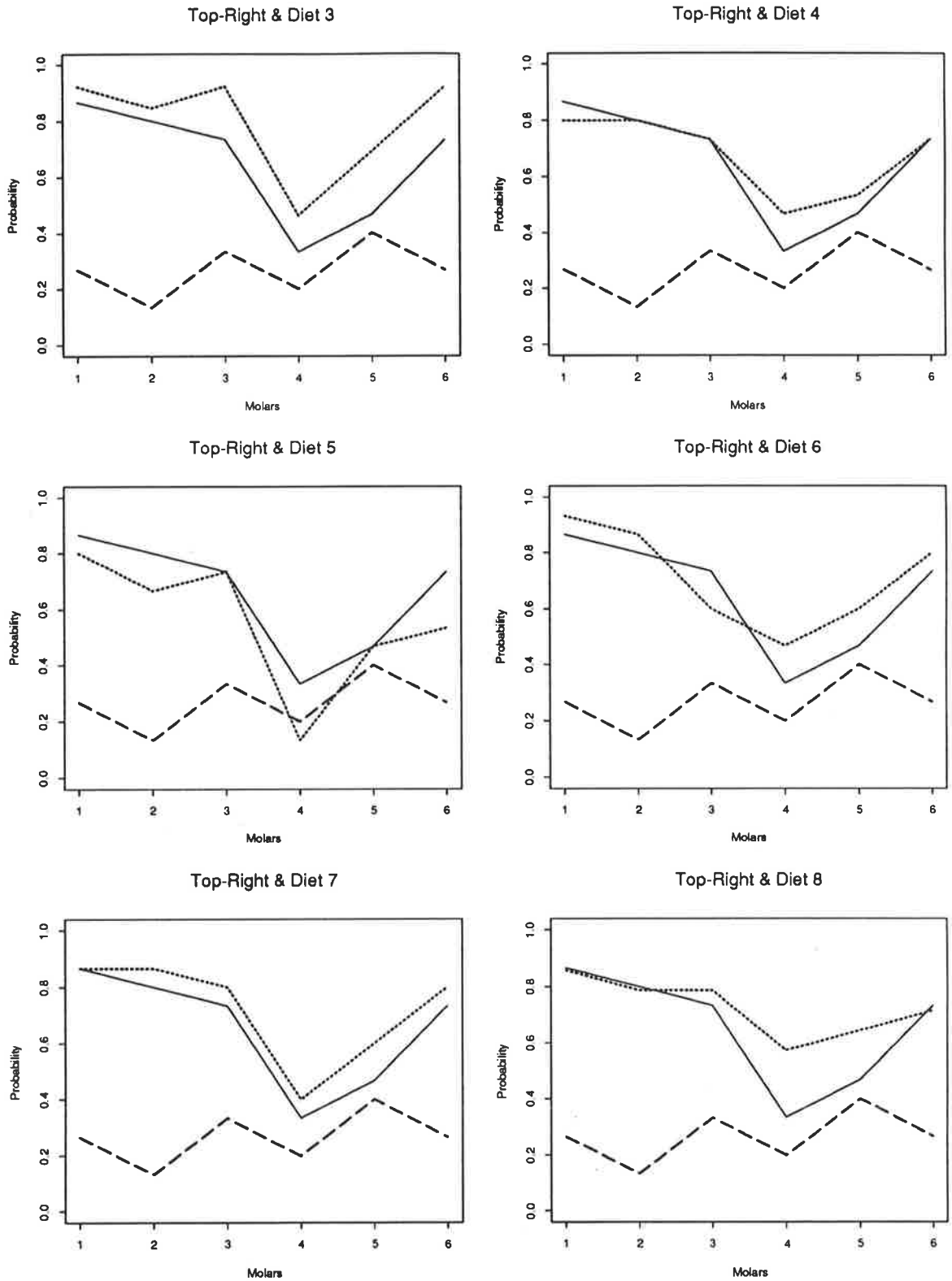

Figure 8.5: Observed probabilities for top-right portion of the mouth

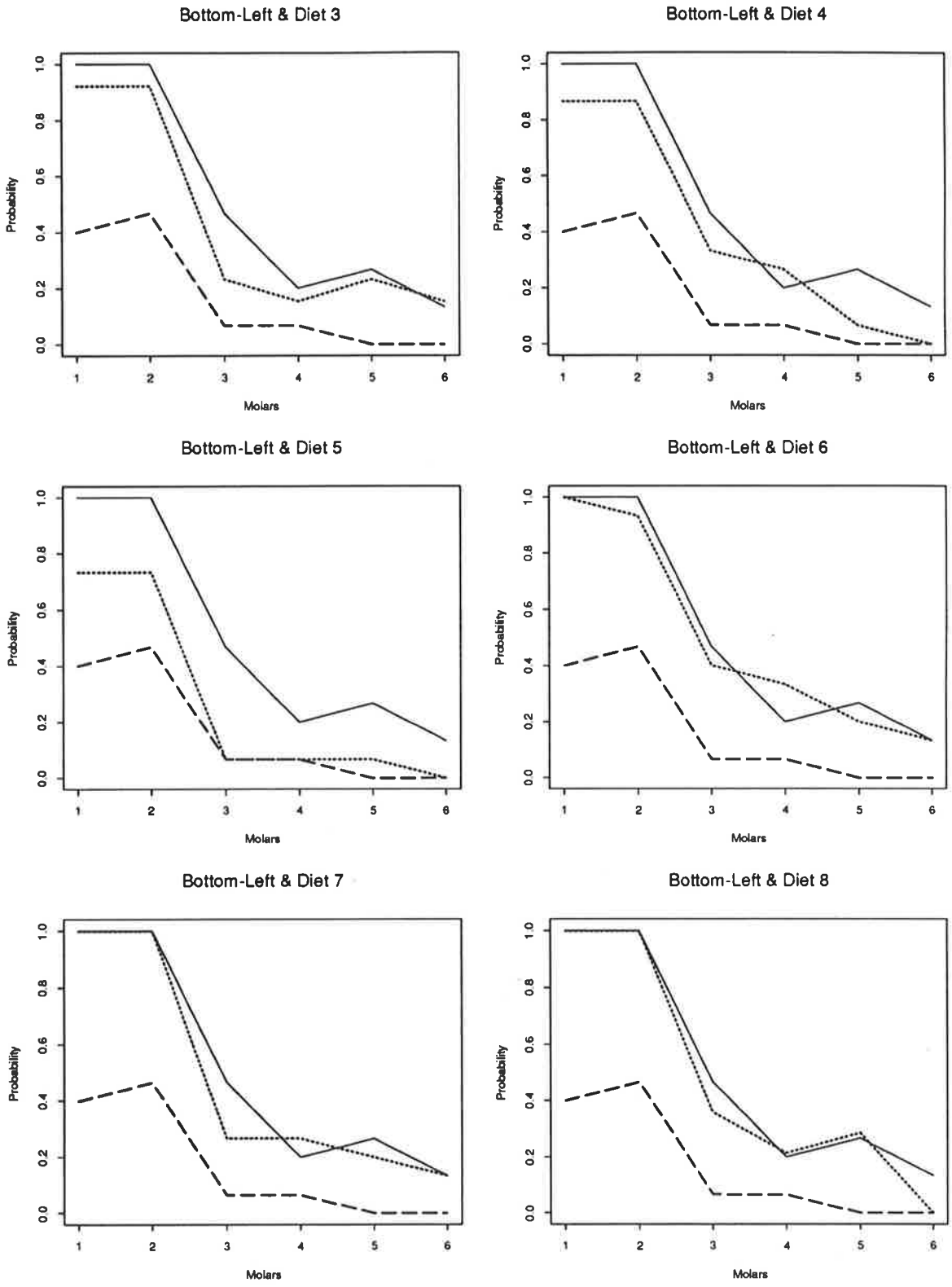

Figure 8.6: Observed probabilities for bottom-left portion of the mouth

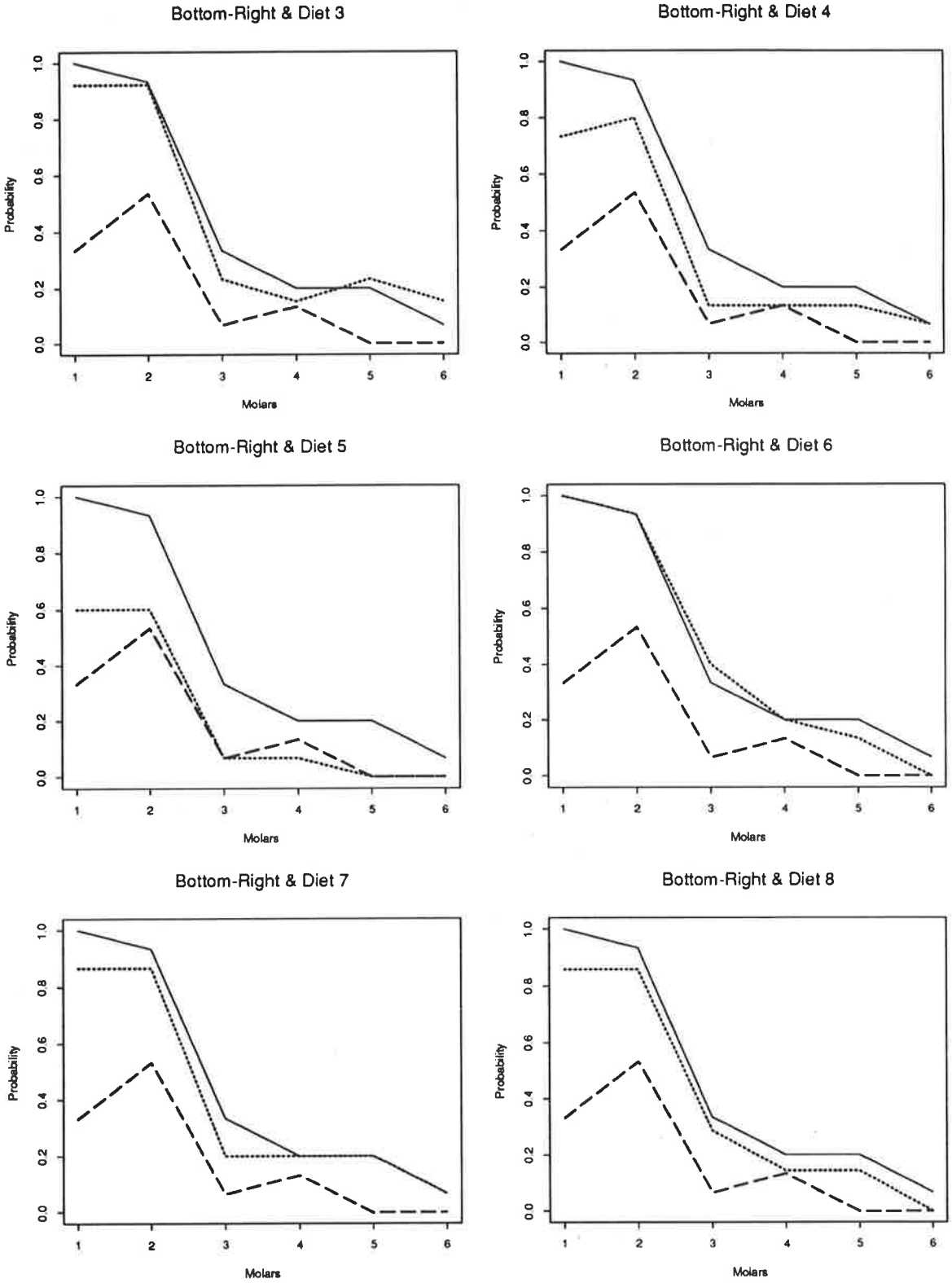

Figure 8.7: Observed probabilities for bottom-right portion of the mouth

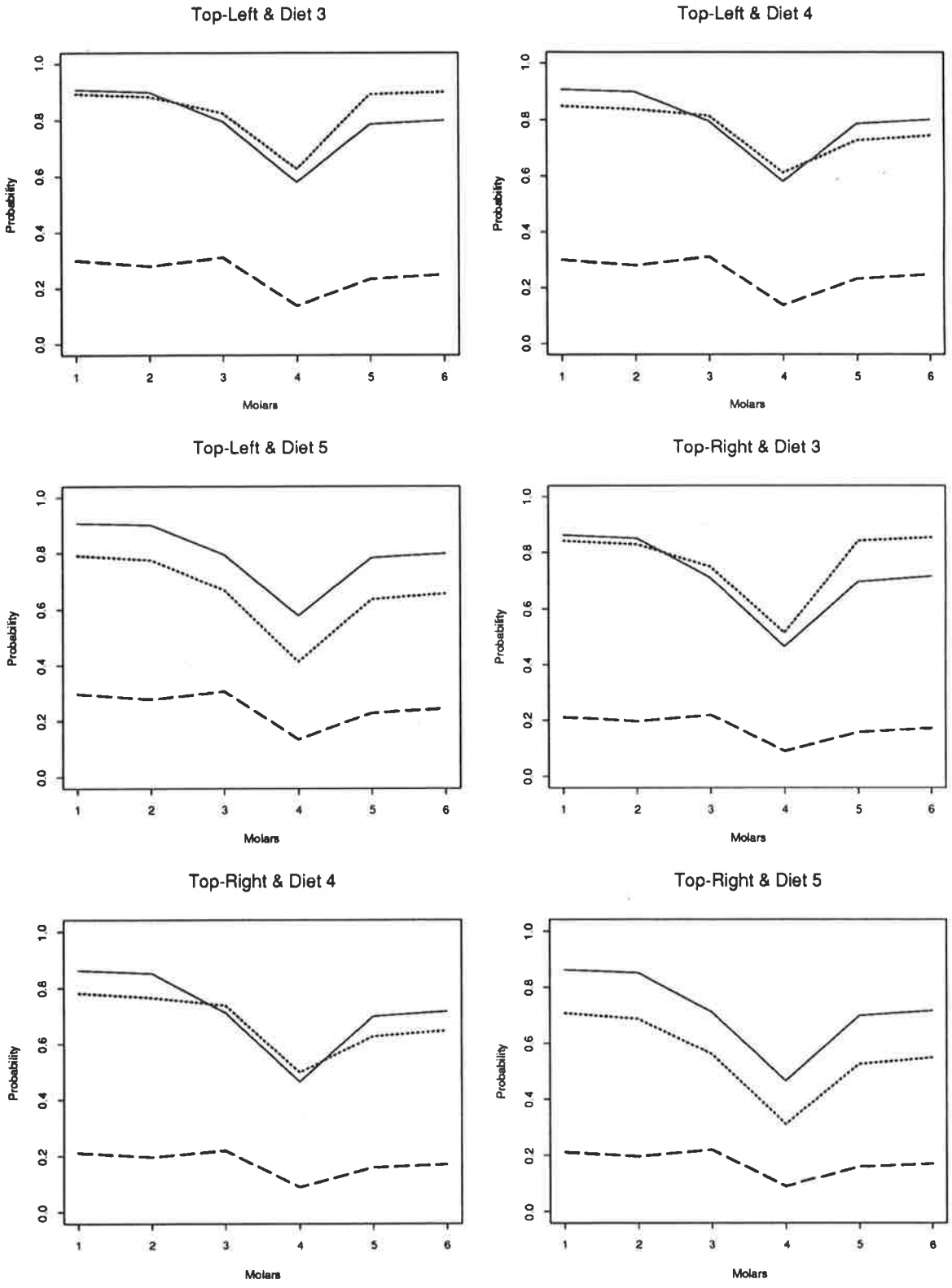

Figure 8.8: Extended mean model under Model 1 and top jaw

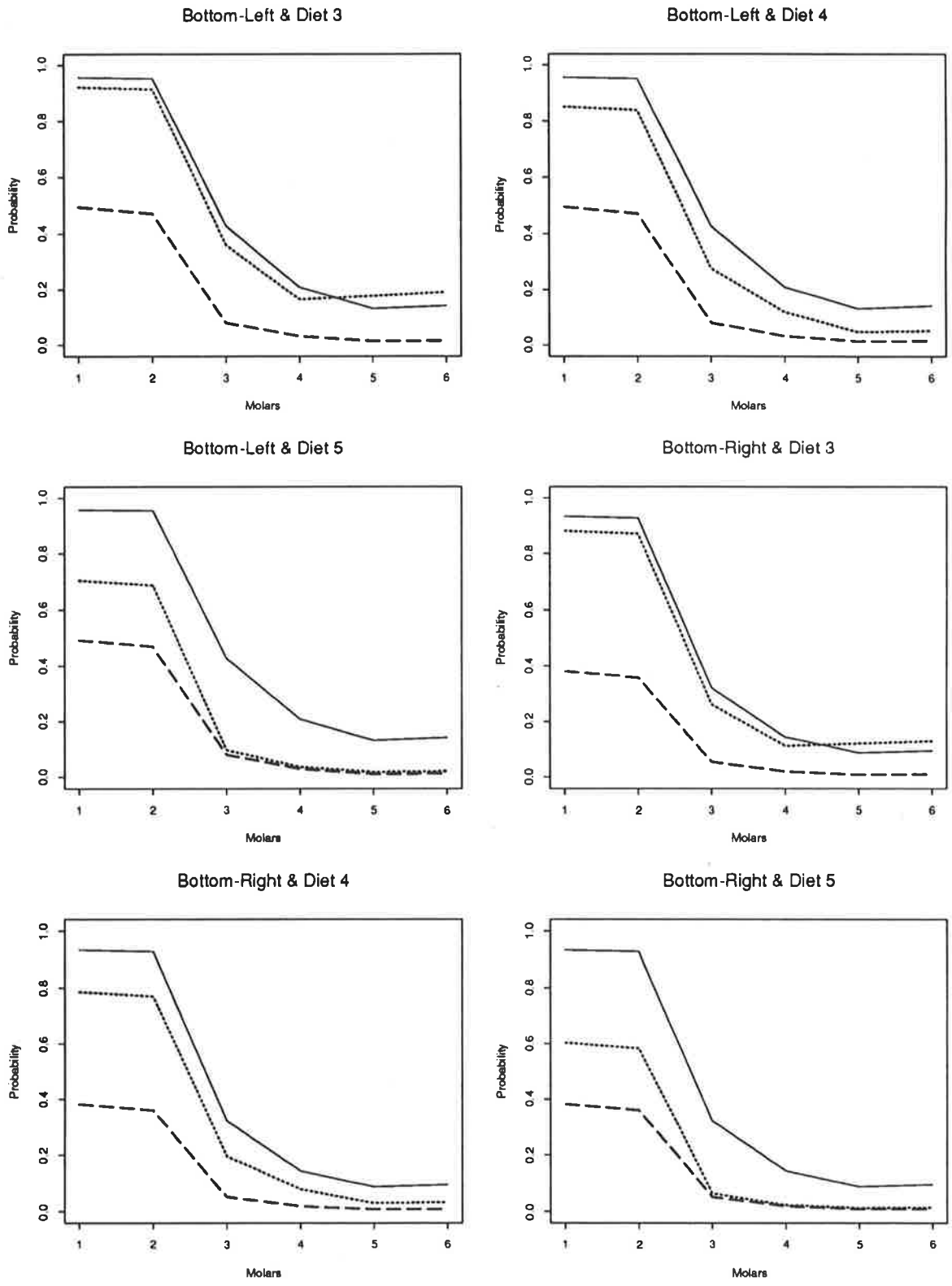

Figure 8.9: Extended mean model under Model 1 and bottom jaw

Appendix A

We can approximate $n^{1/2}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) = [n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})',\ n^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})',\ n^{1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})']'$ using a Taylor expansion about $\boldsymbol{\lambda}$ to equal

$$[-n^{-1}\sum_{i=1}^{n}\delta U_i\{\boldsymbol{\lambda}, \hat{\boldsymbol{\xi}}(\boldsymbol{\lambda})\}/\delta\boldsymbol{\lambda}]^{-1}[n^{-1/2}\sum_{i=1}^{n}U_i(\boldsymbol{\lambda}, \hat{\boldsymbol{\xi}}(\boldsymbol{\lambda}))] \tag{A.1}$$

where

$$U_i(\boldsymbol{\lambda}, \hat{\boldsymbol{\xi}}) = \sum_{i=1}^{n}\boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{S}_i$$

and

$$\frac{\delta U_i(\boldsymbol{\lambda}, \hat{\boldsymbol{\xi}})}{\delta\boldsymbol{\lambda}} = \frac{\partial U_i(\boldsymbol{\lambda}, \hat{\boldsymbol{\xi}})}{\partial\boldsymbol{\lambda}} + \frac{\partial U_i(\boldsymbol{\lambda}, \hat{\boldsymbol{\xi}})}{\partial\hat{\boldsymbol{\xi}}}\frac{\partial\hat{\boldsymbol{\xi}}}{\partial\boldsymbol{\lambda}}.$$

A Taylor expansion about $\boldsymbol{\xi}$, $\boldsymbol{\lambda}$ fixed, gives

$$\begin{aligned} n^{-1/2}\sum_{i=1}^{n}U_i(\boldsymbol{\lambda}, \hat{\boldsymbol{\xi}}(\boldsymbol{\lambda})) &= n^{-1/2}\sum_{i=1}^{n}U_i(\boldsymbol{\lambda}, \boldsymbol{\xi}) + \left\{n^{-1}\sum_{i=1}^{n}\partial U_i(\boldsymbol{\lambda}, \boldsymbol{\xi})/\partial\boldsymbol{\xi}\right\}n^{1/2}(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}) + o_p(1) \\ &= n^{-1/2}\sum_{i=1}^{n}U_i(\boldsymbol{\lambda}, \boldsymbol{\xi}) + o_p(1). \end{aligned} \tag{A.2}$$

This follows from $\hat{\boldsymbol{\xi}}(\boldsymbol{\lambda})$ being $n^{1/2}$-consistent, that is $n^{1/2}(\hat{\boldsymbol{\xi}}(\boldsymbol{\lambda}) - \boldsymbol{\xi})$ is $o_p(1)$, and because $\partial U_i(\boldsymbol{\lambda},\ \boldsymbol{\xi})/\partial\boldsymbol{\xi}$ is a linear function of the $\boldsymbol{S}_i$'s which have mean zero, then

$$n^{-1}\sum_{i=1}^{n}\partial U_i(\boldsymbol{\lambda}, \boldsymbol{\xi})/\partial\boldsymbol{\xi}$$

is $o_p(1)$. The asymptotic multivariate distribution of $n^{-1/2}\sum_{i=1}^{n}U_i(\boldsymbol{\lambda}, \hat{\boldsymbol{\xi}}(\boldsymbol{\lambda}))$ is thus

$$N\left(\mathbf{0},\ n^{-1}\sum_{i=1}^{n}\boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\mathrm{cov}(\boldsymbol{f}_i)\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right).$$

Since $n^{-1}\sum_{i=1}^{n}\partial U_i(\boldsymbol{\lambda},\hat{\boldsymbol{\xi}}(\boldsymbol{\lambda}))/\partial\hat{\boldsymbol{\xi}}$ and $\partial\hat{\boldsymbol{\xi}}/\partial\boldsymbol{\lambda}$ are both $o_p(1)$ then

$$-n^{-1}\sum_{i=1}^{n}\delta U_i\{\boldsymbol{\lambda},\hat{\boldsymbol{\xi}}(\boldsymbol{\lambda})\}/\delta\boldsymbol{\lambda} \tag{A.3}$$

converges in probability to the limit of

$$-n^{-1}\sum_{i=1}^{n}\partial U_i\{\boldsymbol{\lambda},\hat{\boldsymbol{\xi}}(\boldsymbol{\lambda})\}/\partial\boldsymbol{\lambda}.$$

Finally, by (A.2), we have

$$-n^{-1}\sum_{i=1}^{n}\partial U_i\{\boldsymbol{\lambda},\hat{\boldsymbol{\xi}}(\boldsymbol{\lambda})\}/\partial\boldsymbol{\lambda} = -n^{-1}\sum_{i=1}^{n}\boldsymbol{D}_i'\boldsymbol{V}_i^{-1}(\boldsymbol{D}_i-\partial\boldsymbol{f}_i/\partial\boldsymbol{\lambda})+o_p(1), \tag{A.4}$$

where

$$\partial\boldsymbol{f}_i/\partial\boldsymbol{\lambda} = \begin{bmatrix} 0 & 0 & 0 \\ \partial\boldsymbol{d}_i/\partial\boldsymbol{\beta} & 0 & 0 \\ \partial\boldsymbol{r}_i/\partial\boldsymbol{\beta} & \partial\boldsymbol{r}_i/\partial\boldsymbol{\gamma} & 0 \end{bmatrix}.$$

Consequently as (A.3) converges in probability to the first term on the right-side of (A.4), then the Normality of $n^{1/2}(\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda})$ follows. The asymptotic variance of $n^{1/2}(\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda})$ using (A.1) leads to (6.8).

Bibliography

[1] A. Agresti. A survey of models for repeated ordered categorical response data. *Statistics in Medicine*, 8:1209–1224, 1989.

[2] A. C. Aitken. *Determinants and Matrices.* Oliver and Boyd, London, 9th edition, 1956.

[3] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, pages 267–281, Budapest: Akademia Kiado, 1973.

[4] D. F. Andrews and A. M. Herzberg. *Data: A Collection of Problems from Many Fields for the Student and Research Worker.* Springer-Verlag, New York, 1985.

[5] D. F. Archer. Prolactin response to thyrotropin-releasing hormone in women with infertility and/or randomly elevated serum prolactin levels. *Fertility And Sterility*, 47:559–564, 1987.

[6] R. B. Avery, L. P. Hansen, and V. J. Hotz. Multiperiod probit models and orthogonality condition estimation. *International Economic Review*, 24:21–35, 1983.

[7] J. K. Baksalary, L. C. A. Corsten, and R. Kala. Reconciliation of two different views on estimation of growth curve parameters. *Biometrika*, 65:662–665, 1978.

[8] P. J. Beitler and J. R. Landis. A mixed-effects model for categorical data. *Biometrics*, 41:991–1000, 1985.

[9] D. A. Binder. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51:279–292, 1983.

[10] Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete Multivariate Analysis: Theory and Practice.* The MIT Press, U. S. A., 1975.

[11] L. Boltzmann. Über die Beziehung zwischen dem zweitin Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung repective den Sätzen über das Wärmegleichgewicht. *Wiener Berichte*, 76:373–435, 1877.

[12] D. G. Bonett, J. A. Woodward, and P. M. Bentler. Some extensions of a linear model for categorical variables. *Biometrics*, 41:745–750, 1985.

[13] G. E. P. Box. Problems in the analysis of growth and wear curves. *Biometrics*, 6:363–389, 1950.

[14] H. Bozdogan. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika*, 52:345–370, 1987.

[15] N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25, 1993.

[16] C. J. Brien, W. N. Venables, A. T. James, and O. Mayo. An analysis of correlation matrices: equal correlations. *Biometrika*, 71:545–554, 1984.

[17] M. A. Cameron and G. K. Eagleson. A new procedure for assessing large sets of correlations. *Austral. J. Statist.*, 27:84–95, 1985.

[18] V. Carey, S. L. Zeger, and P. Diggle. Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, 80:517–526, 1993.

[19] G. J. Carr and E. M. Chi. Analysis of variance for repeated measures data: a generalized estimating equations approach. *Statistics in Medicine*, 11:1033–1040, 1992.

[20] R. J. Carroll and D. Ruppert. Robust estimation in heteroscedastic linear models. *The Annals of Statistics*, 10:429–441, 1982.

[21] P. J. Catalano and L. M. Ryan. Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*, 87:651–658, 1992.

[22] E. M. Chi and G. C. Reinsel. Models for longitudinal data with random effects and AR(1) errors. *Journal of the American Statistical Association*, 84:452–459, 1989.

[23] V. M. Chinchilli and W. H. Carter. A likelihood ratio test for a patterned covariance matrix in a multivariate growth-curve model. *Biometrics*, 40:151–156, 1984.

[24] J. B. Cologne, R. L. Carter, S. Fujita, and S. Ban. Application of generalized estimating equations to a study of *in vitro* radiation sensitivity. *Biometrics*, 49:927–934, 1993.

[25] D. M. Cooper and R. Thompson. A note on the estimation of the parameters of the autoregressive-moving average process. *Biometrika*, 64:625–628, 1977.

[26] D. R. Cox and N. Reid. Parameter orthogonality and approximate conditional inference. *J. R. Statist. Soc. B*, 49:1–39, 1987.

[27] B. R. Cullis and C. A. McGilchrist. A model for the analysis of growth data from designed experiments. *Biometrics*, 46:131–142, 1990.

[28] B. R. Cullis and A. P. Verbyla. Nonlinear regression modelling and time dependent covariates in repeated measures experiments. *Austral. J. Statist.*, 34:145–160, 1992.

[29] C. V. Damaraju. Category-specific covariate effects in longitudinal study designs. *Commun. Statist.-Theory Meth.*, 22:1441–1469, 1993.

[30] G. A. Darlington and V. T. Farewell. Binary longitudinal data analysis with correlation a function of explanatory variables. *Biom. J.*, 34:899–910, 1992.

[31] M. Davidian and R. J. Carroll. Variance function estimation. *Journal of the American Statistical Association*, 82:1079–1091, 1987.

[32] A. P. Dempster, D. B. Rubin, and R. K. Tsutakawa. Estimation in covariance component models. *Journal of the American Statistical Association*, 76:341–353, 1981.

[33] M. L. Drum and P. McCullagh. REML estimation with exact covariance in the logistic mixed model. *Biometrics*, 49:677–689, 1993.

[34] Q. P. Duong. On the choice of the order of autoregressive models: a ranking and selection approach. *J. Time Ser. Analysis*, 5:145–157, 1984.

[35] G. K. Eagleson. A robust test for multiple comparisons of correlation coefficients. *Austral. J. Statist.*, 25:256–263, 1983.

[36] R. C. Elston. On estimating time response curves. *Biometrics*, 20:643–647, 1964.

[37] R. C. Elston and J. E. Grizzle. Estimation of time-response curves and their confidence bands. *Biometrics*, 18:148–159, 1962.

[38] J. C. Evans and E. A. Roberts. Analysis of sequential observations with applications to experiments on grazing animals and perennial plants. *Biometrics*, 35:687–693, 1979.

[39] E. J. Farris and J. Q. Griffith. *The Rat in Laboratory Investigation.* J. B. Lippincott Company, U. S. A., 2nd edition, 1949.

[40] T. Fearn. A two-stage model for growth curves which leads to Rao's covariance adjusted estimators. *Biometrika*, 64:141–143, 1977.

[41] D. Firth. On the efficiency of quasi-likelihood estimation. *Biometrika*, 74:233–245, 1987.

[42] D. Firth. *Statistical Theory and Modelling: In honour of Sir David Cox, FRS*, pages 55–82. D. V. Hinkley, N. Reid & E. J. Snell (ed.). Chapman & Hall, London, 1991.

[43] G. M. Fitzmaurice and N. M. Laird. A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, 80:141–151, 1993.

[44] G. M. Fitzmaurice, N. M. Laird, and A. G. Rotnitzky. Regression models for discrete longitudinal responses. *Statistical Science*, 8:284–309, 1993.

[45] D. A. Follmann. Growth curve models with restrictions on random parameters. *Commun. Statist.-Theory Meth.*, 21:2775–2795, 1992.

[46] R. N. Forthofer and G. G. Koch. An analysis for compounded functions of categorical data. *Biometrics*, 29:143–157, 1973.

[47] Y. Fujikoshi and C. R. Rao. Selection of covariables in the growth curve model. *Biometrika*, 78:779–785, 1991.

[48] L. J. Gleser and I. Olkin. *Linear Models in Multivariate Analysis*, pages 267–292. Essays in Probability and Statistics, R. C. Bose *et al.* (ed.). Wiley, New York, 1970.

[49] V. P. Godambe and C. C. Heyde. Quasi-likelihood and optimal estimation. *International Statistical Review*, 55:231–244, 1987.

[50] C. Gourieroux, A. Monfort, and A. Trognon. Pseudo maximum likelihood methods: theory. *Econometrica*, 52:681–700, 1984.

[51] J. E. Grizzle and D. Allen. Analysis of growth and dose response curves. *Biometrics*, 25:357–381, 1969.

[52] J. E. Grizzle, C. F. Starmer, and G. G. Koch. Analysis of categorical data by linear models. *Biometrics*, 25:489–504, 1969.

[53] D. A. Harville. Bayesian inference for variance components using only error contrasts. *Biometrika*, 61:383–385, 1974.

[54] D. A. Harville. Extension of the Gauss-Markov theorem to include the estimation of random effects. *The Annals of Statistics*, 76:384–395, 1976.

[55] D. A. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72:320–340, 1977.

[56] R. Hebel and M. W. Stromberg. *Anatomy and Embryology of the Laboratory Rat.* Bio. Med. Verlag, Federal Republic of Germany, 1986.

[57] M. Hills. A note on the analysis of growth curves. *Biometrics*, 24:189–196, 1968.

[58] M. Hills. On looking at large correlation matrices. *Biometrika*, 56:249–253, 1969.

[59] R. I. Jennrich and M. D. Schluchter. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42:805–820, 1986.

[60] N. L. Johnson and S. Kotz. *Distributions in Statistics: Continuous Multivariate Distributions*, pages 216–220. Wiley, New York, 1972.

[61] R. H. Jones. *Longitudinal Data with Serial Correlation: A State-space Approach.* Chapman and Hall, London, 1993.

[62] R. L. Kashyap. Optimal choice of AR and MA parts in autoregressive moving average models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4:99–104, 1982.

[63] C. G. Khatri. A note on a MANOVA model applied to problems in growth curve. *Ann. Inst. Statist. Math.*, 18:75–86, 1966.

[64] D. G. Kleinbaum. A generalisation of the growth curve model which allows missing data. *Journal of Multivariate Analysis*, 3:117–124, 1973.

[65] G. G. Koch, P. B. Imrey, and D. W. Reinfurt. Linear model analysis of categorical data with incomplete response vectors. *Biometrics*, 28:663–692, 1972.

[66] G. G. Koch, J. R. Landis, J. L. Freeman, D. H. Freeman, and R. G. Lehnen. A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics*, 33:133–158, 1977.

[67] E. L. Korn and A. S. Whittemore. Methods for analyzing panel studies of acute health effects of air polution. *Biometrics*, 35:795–802, 1979.

[68] T. Kubokawa. Two-stage procedures for parameters in a growth curve model. *Journal of Statistical Planning and Inference*, 22:105–115, 1989.

[69] T. Kubokawa, A. K. Saleh, and K. Morita. Improving on MLE of coefficient matrix in a growth curve model. *Journal of Statistical Planning and Inference*, 31:169–177, 1992.

[70] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[71] N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38:963–974, 1982.

[72] N. Lange and N. M. Laird. The effect of covariance structure on variance estimation in balanced growth-curve models with random parameters. *Journal of the American Statistical Association*, 84:241–247, 1989.

[73] A. Laor and A. Cohen. Analysis of premature ventricular counts in long term ECG following myocardial infarction. *Statistics in Medicine*, 11:963–973, 1992.

[74] M. W. J. Layard. Large sample tests for the equality of two covariance matrices. *Ann. Math. Statist.*, 43:123–141, 1972.

[75] J. C. Lee. Tests and model selection for the general growth curve model. *Biometrics*, 47:147–159, 1991.

[76] Y. H. K. Lee. A note on Rao's reduction of Potthoff and Roy's generalized linear model. *Biometrika*, 61:349–351, 1974.

[77] F. B. Leech and M. J. R. Healy. The analysis of experiments on growth rate. *Biometrics*, 15:98–106, 1959.

[78] M. Lefkopoulou, D. Moore, and L. Ryan. The analysis of multiple correlated binary outcomes: application to rodent teratology experiments. *Journal of the American Statistical Association*, 84:810–815, 1989.

[79] M. Lefkopoulou and L. Ryan. Global tests for multiple binary outcomes. *Biometrics*, 49:975–988, 1993.

[80] I. E. Leppik *et al.* A double-blind crossover evaluation of progabide in partial seizures. *Neurology*, 35:385, 1985.

[81] K. Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22, 1986.

[82] K. Y. Liang, S. L. Zeger, and B. Qaqish. Multivariate regression analyses for categorical data. *J. R. Statist. Soc. B*, 54:3–40, 1992.

[83] M. J. Lindstrom and D. M. Bates. Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83:1014–1022, 1988.

[84] M. J. Lindstrom and D. M. Bates. Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46:673–687, 1990.

[85] S. R. Lipsitz, N. M. Laird, and D. P. Harrington. Finding the design matrix for the marginal homogeneity model. *Biometrika*, 77:353–358, 1990.

[86] S. R. Lipsitz, N. M. Laird, and D. P. Harrington. Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika*, 78:153–160, 1991.

[87] S. R. Lipsitz, N. M. Laird, and D. P. Harrington. A three-stage estimator for studies with repeated and possibly missing binary outcomes. *Appl. Statist.*, 41:203–213, 1992.

[88] E. P. Liski. Detecting influential measurements in a growth curves model. *Biometrics*, 47:659–668, 1991.

[89] S. Lundbye-Christensen. A multivariate growth curve model for pregnancy. *Biometrics*, 47:637–657, 1991.

[90] L. Mancl. *Regression analysis of correlated discrete and continuous data: evaluation of an estimating equation approach.* Ph. D. dissertation, Dept. Biostatistics, Univ. Washington, 1992.

[91] P. McCullagh. Regression models for ordinal data. *J. R. Statist. Soc. B*, 42:109–142, 1980.

[92] P. McCullagh and J. A. Nelder. *Generalized Linear Models.* Chapman and Hall, London, 2nd edition, 1989.

[93] R. D. Mensah, R. K. Elswick, and V. M. Chinchilli. Consistent estimators of the variance-covariance matrix of the GMANOVA model with missing data. *Commun. Statist.-Theory Meth.*, 22:1495–1514, 1993.

[94] M. E. Miller, C. S. Davis, and J. R. Landis. The analysis of longitudinal polytomous data: generalized estimating equations and connections with weighted least squares. *Biometrics*, 49:1033–1044, 1993.

[95] M. E. Miller and J. R. Landis. Generalized variance component models for clustered categorical response variables. *Biometrics*, 47:33–44, 1991.

[96] P. A. P. Moran. Testing the largest of a set of correlation coefficients. *Austral. J. Statist.*, 22:289–297, 1980.

[97] D. F. Morrison. *Multivariate Statistical Methods.* McGraw-Hill, New York, 1967.

[98] R. J. Muirhead. *Aspects of Multivariate Statistical Analysis.* Wiley, New York, 1982.

[99] J. A. Nelder and Y. Lee. Likelihood, quasi-likelihood and pseudolikelihood: some comparisons. *J. R. Statist. Soc. B*, 54:273–284, 1992.

[100] J. A. Nelder and D. Pregibon. An extended quasi-likelihood function. *Biometrika*, 74:221–232, 1987.

[101] J. Ogawa and G. Ishii. The relationship algebra and the analysis of variance of a partially balanced incomplete block design. *Ann. Math. Statist.*, 36:1815–1828, 1965.

[102] R. J. O'Hara Hines and J. F. Lawless. Modelling overdispersion in toxicological mortality data grouped over time. *Biometrics*, 49:107–121, 1993.

[103] I. Olkin. *Inference for a normal population when the parameters exhibit some structure*, pages 759–773. Reliability and Biometry: Statistical Analysis of Life-length, F. Proschan and R. J. Serfling (ed.). Society for Industrial and Applied Mathematics, Philadelphia, 1974.

[104] M. C. Paik. Parametric variance function estimation for nonnormal repeated measurement data. *Biometrics*, 48:19–30, 1992.

[105] T. Park and C. S. Davis. A test of the missing data mechanism for repeated categorical data. *Biometrics*, 49:631–638, 1993.

[106] T. Park, S. Lee, and R. F. Woolson. A test of the missing data mechanism for repeated measures data. *Commun. Statist.-Theory Meth.*, 22:2813–2829, 1993.

[107] H. D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58:545–554, 1971.

[108] R. F. Potthoff and S. N. Roy. A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51:313–326, 1964.

[109] R. L. Prentice. Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 44:1033–1048, 1988.

[110] R. L. Prentice and L. P. Zhao. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, 47:825–839, 1991.

[111] B. F. Qaqish and K. Y. Liang. Marginal models for correlated binary responses with multiple classes and multiple levels of nesting. *Biometrics*, 48:939–950, 1992.

[112] C. R. Rao. Some problems involving linear hypotheses in multivariate analysis. *Biometrika*, 46:49–58, 1959.

[113] C. R. Rao. The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, 52:447–458, 1965.

[114] C. R. Rao. *Covariance Adjustment and Related Problems in Multivariate Analysis*, pages 87–103. Multivariate Analysis, P. R. Krishnaiah (ed.). Academic Press, New York, 1966.

[115] C. R. Rao. *Least Squares Theory using an estimated dispersion matrix and its application to measurement of signals*, pages 355–372. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, C. R. Rao (ed.). volume 1, Berkeley, 1967.

[116] G. Reinsel. Multivariate repeated-measurement or growth curve models with multivariate random-effects covariance structure. *Journal of the American Statistical Association*, 77:190–195, 1982.

[117] G. Reinsel. Estimation and prediction in a multivariate random effects generalized linear model. *Journal of the American Statistical Association*, 79:406–414, 1984.

[118] P. J. Ricci and A. P. Verbyla. Generalized estimating equations and the sum of profiles model. Research Report 91/6, The University of Adelaide, Department of Statistics, 1991.

[119] P. J. Ricci and A. P. Verbyla. Regression models of repeated binary response. Research Report 91/1, The University of Adelaide, Department of Statistics, 1991.

[120] P. J. Ricci, A. P. Verbyla, and W. N. Venables. Residual maximum likelihood and the growth curve model. *Austral. J. Statist.*, submitted.

[121] A. Rotnitzky and N. P. Jewell. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, 77:485–497, 1990.

[122] J. G. Rowell and D. E. Walters. Analysing data with repeated observations on each experimental unit. *The Journal of Agricultural Science*, 87:423–432, 1976.

[123] H. G. Q. Rowett. *The Rat as a small Mammal.* John Murray, London, 1957.

[124] D. B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

[125] R. L. Sandland and C. A. McGilchrist. Stochastic growth curve analysis. *Biometrics*, 35:255–271, 1979.

[126] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[127] T. Schweder and E. Spjøtvoll. Plots of P-values to evaluate many tests simultaneously. *Biometrika*, 69:493–502, 1982.

[128] S. R. Searle. *Linear Models.* Wiley, New York, 1971.

[129] G. A. F. Seber. *Multivariate Observations.* Wiley, New York, 1984.

[130] P. J. Smith and A. Hadgu. Sensitivity and specificity for correlated observations. *Statistics in Medicine*, 11:1503–1509, 1992.

[131] G. K. Smyth. Generalized linear models with varying dispersion. *J. R. Statist. Soc. B*, 51:47–60, 1989.

[132] J. N. Srivastava and L. L. McDonald. Analysis of growth curves under the hierarchical models. *Sankhya A*, 36:251–260, 1974.

[133] M. S. Srivastava. Multivariate data with missing observations. *Commun. Statist.-Theory Meth.*, 14:775–792, 1985.

[134] E. J. Stanek III and S. R. Diehl. Growth curve models of repeated binary response. *Biometrics*, 44:973–983, 1988.

[135] W. M. Stanish, D. B. Gillings, and G. G. Koch. An application of multivariate ratio methods for the analysis of a longitudinal clinical trial with missing data. *Biometrics*, 34:305–317, 1978.

[136] R. Stiratelli, N. Laird, and J. H. Ware. Random-effects models for serial observations with binary response. *Biometrics*, 40:961–971, 1984.

[137] D. O. Stram, L. J. Wei, and J. H. Ware. Analysis of repeated ordered categorical outcomes with possibly missing observations and time-dependent covariates. *Journal of the American Statistical Association*, 83:631–637, 1988.

[138] T. A. Stukel. Comparison of methods for the analysis of longitudinal interval count data. *Statistics in Medicine*, 12:1339–1351, 1993.

[139] N. Sugiura and T. Kubokawa. Estimating common parameters of growth curve models. *Ann. Inst. Statist. Math.*, 40:119–135, 1988.

[140] W. H. Swallow and J. F. Monahan. Monte Carlo comparison of ANOVA, MIVQUE, REML, and ML estimators of variance components. *Technometrics*, 26:47–57, 1984.

[141] P. F. Thall. Mixed poisson likelihood regression models for longitudinal interval count data. *Biometrics*, 44:197–209, 1988.

[142] P. F. Thall and S. C. Vail. Some covariance models for longitudinal count data with overdispersion. *Biometrics*, 46:657–671, 1990.

[143] K. T. Tsai and J. A. Koziol. Score and Wald tests for the multivariate growth curve model with missing data. *Ann. Inst. Statist. Math.*, 40:179–186, 1988.

[144] K. T. Tsai and J. A. Koziol. Score and Wald tests for the multivariate growth curve model with missing data and a patterned covariance matrix. *Commun. Statist.-Theory Meth.*, 22:311–317, 1993.

[145] A. P. Verbyla. A note on the inverse covariance matrix of the autoregressive process. *Austral. J. Statist.*, 27:221–224, 1985.

[146] A. P. Verbyla. Conditioning in the growth curve model. *Biometrika*, 73:475–483, 1986.

[147] A. P. Verbyla. *Extensions to profile analysis.* Unpublished Ph. D. Thesis, Department of Statistics, The University of Adelaide, 1986.

[148] A. P. Verbyla. A conditional derivation of residual maximum likelihood. *Austral. J. Statist.*, 32:227–230, 1990.

[149] A. P. Verbyla and B. R. Cullis. Modelling in repeated measures experiments. *Appl. Statist.*, 39:341–356, 1990.

[150] A. P. Verbyla and B. R. Cullis. The analysis of multistratum and spatially correlated repeated measures data. *Biometrics*, 48:1015–1032, 1992.

[151] A. P. Verbyla and W. N. Venables. An extension of the growth curve model. *Biometrika*, 75:129–138, 1988.

[152] D. Von Rosen. Maximum likelihood estimators in multivariate linear model. *Journal of Multivariate Analysis*, 31:187–200, 1989.

[153] D. Von Rosen. Moments for a multivariate linear model with an application to the growth curve model. *Journal of Multivariate Analysis*, 35:243–259, 1990.

[154] D. Von Rosen. Moments of maximum likelihood estimators in the growth curve model. *Statistics*, 22:111–131, 1991.

[155] M. Von Tress. Longitudinal models for polytomous responses. *Commun. Statist.-Theory Meth.*, 22:3523–3536, 1993.

[156] J. H. Ware. Linear models for the analysis of longitudinal studies. *The American Statistician*, 39:95–101, 1985.

[157] J. H. Ware, S. Lipsitz, and F. E. Speizer. Issues in the analysis of repeated categorical outcomes. *Statistics in Medicine*, 7:95–107, 1988.

[158] R. W. M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61:439–447, 1974.

[159] L. J. Wei and D. O. Stram. Analysing repeated measurements with possibly missing observations by modelling marginal distributions. *Statistics in Medicine*, 7:139–148, 1988.

[160] S. Weisberg. *Applied Linear Regression.* Wiley, New York, second edition, 1980.

[161] J. Wishart. Growth rate determination in nutrition studies with the bacon pig, and their analysis. *Biometrika*, 30:16–28, 1938.

[162] R. F. Woolson and W. R. Clarke. Analysis of categorical incomplete longitudinal data. *J. R. Statist. Soc. A*, 147:87–99, 1984.

[163] R. F. Woolson and J. D. Leeper. Growth curve analysis of complete and incomplete longitudinal data. *Commun. Statist.-Theory Meth.*, A9:1491–1513, 1980.

[164] R. F. Woolson, J. D. Leeper, and W. R. Clarke. Analysis of incomplete data from longitudinal and mixed longitudinal studies. *J. R. Statist. Soc. A*, 141:242–252, 1978.

[165] D. Wypij, M. Pugh, and J. H. Ware. Modeling pulmonary function growth with regression splines. *Statistica Sinica*, 3:329–350, 1993.

[166] S. L. Zeger. Commentary. *Statistics in Medicine*, 7:161–168, 1988.

[167] S. L. Zeger and K. Y. Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42:121–130, 1986.

[168] S. L. Zeger, K. Y. Liang, and P. S. Albert. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44:1049–1060, 1988.

[169] S. L. Zeger, K. Y. Liang, and S. G. Self. The analysis of binary longitudinal data with time-independent covariates. *Biometrika*, 72:31–38, 1985.

[170] I. Žežula. Covariance components estimation in the growth curve model. *Statistics*, 24:321–330, 1993.

[171] L. P. Zhao and R. L. Prentice. Correlated binary regression using a quadratic exponential model. *Biometrika*, 77:642–648, 1990.

[172] L. P. Zhao, R. L. Prentice, and S. G. Self. Multivariate mean parameter estimation by using a partly exponential model. *J. R. Statist. Soc. B*, 54:805–811, 1992.